UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

DESIGN AND EVALUATION OF ANALYTICAL TOOLS FOR
EMERGENCY DEPARTMENT MANAGEMENT BASED ON MACHINE
LEARNING TECHNIQUES

TESIS DOCTORAL

Alexander Zlotnik Enaliev
Ingeniero de Telecomunicación

2016

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

DESIGN AND EVALUATION OF ANALYTICAL TOOLS FOR
EMERGENCY DEPARTMENT MANAGEMENT BASED ON MACHINE
LEARNING TECHNIQUES

TESIS DOCTORAL

2016

Autor:

Alexander Zlotnik Enaliev
Ingeniero de Telecomunicación

Co-director:

Juan Manuel Montero Martínez
Doctor Ingeniero de Telecomunicación

Co-directora:

Ascensión Gallardo Antolín
Doctora Ingeniera de Telecomunicación

This Ph.D. thesis is dedicated to my parents.

Institutional acknowledgments

The following institutions made this work possible:

- Department of Electronic Engineering of the Technical University of Madrid.
- Ramón y Cajal University Hospital.
- Ramón y Cajal Clinical Research Institute (IRYCIS).

Abstract

The Spanish National Healthcare System (NHS) is mostly publicly funded and provided. It is considered highly cost-efficient according to international studies based on World Health Organization (WHO) data. However, the contention of healthcare costs increases while maintaining adequate levels of quality of care, is still a largely unsolved problem. In recent years, Emergency Departments (EDs) of specialized care hospitals have been subjected to budget restrictions, increased visits and increased clinical complexity of these visits. These circumstances require new approaches to ED management, which could benefit from decision support tools.

In this Ph.D. thesis, we propose machine learning solutions for two problems common to most EDs of specialized care hospitals: ED census forecasting and real-time prediction of probabilities of inpatient admission for all triaged patients in the ED. These solutions could be used as decision support systems. Data for the development of these solutions were provided by the Ramón y Cajal University Hospital of Madrid, a large specialized care referral center with all medical specialties excepting Obstetrics. In years 2011 and 2012 it had approximately 1,100 beds and approximately 553,000 patients assigned to its clinical area. Another topic of this Ph.D. thesis are software tools for the generation of logistic regression and Cox regression nomograms, since nomograms can be used as clinical decision aids and as contingency procedures in case of failure of computer-based decision support systems.

The first topic of this Ph.D. thesis is the development of models for ED census forecasting (i.e. prediction of the number of patients present at the ED at a given time). One of the uses of ED census forecasting is nursing personnel allocation, based on national and international recommendations. We chose an 8-hour granularity for our forecasts since many resources (such as nursing personnel) in the ED are organized in 8-hour shifts. Our aim was to generate forecasts for two dependent variables: average ED census levels and maximum ED census levels. Maximum ED census forecasts within 8-hour shifts could be used for nursing personnel allocation, while average ED census forecasts within 8-hour shifts could be used for the other needs (such as allocation of administrative personnel).

We used a generalized regression approach to time series forecasting with several machine learning algorithms: M5P, Alternating Model Trees (AMT) and Support Vector Regression (SVR). We compared these to a series of benchmarks: usual nursing staffing levels (and usual resource allocation policies), stratified average (averages stratified by the three 8-

hour shifts of a day), linear regression and Seasonal Autoregressive Integrated Moving Average (SARIMA) models. Forecasts were generated for both dependent variables: average ED census levels and maximum ED census levels. Four forecast horizons were tested: 1 week, 2 weeks, 4 weeks and 8 weeks. Underestimation risks, overestimation risks and approximations to monetary costs of resource allocations policies were defined for both average and maximum ED census forecasts. Maximum ED census forecasts were transformed into nursing personnel levels, and underestimation and overestimation risks for maximum ED census forecasts were transformed into understaffing and overstaffing risks. A custom training and evaluation scheme was used, with increasingly larger train sets and fixed-length test sets. The same scheme but with fixed-length train sets of 1 year and fixed-length test sets was also used. The latter approach did not improve the results.

In the case of *maximum* ED census forecasts, M5P was the best choice for the reduction of major and medium nursing personnel understaffing risks, for all forecast horizons. Compared to the *usual staffing levels*, personnel planning with M5P could reduce major understaffing (>9 nurse) risks more than 10-fold (a reduction to ~1% with M5P compared to ~13% obtained with the usual nurse staffing levels); and could reduce medium understaffing (7-9 nurses) risks approximately 3-fold (a reduction to ~3% with M5P compared to ~10% obtained with the usual staffing levels). The usage of M5P also implied 5% - 6.1% increases in estimated nursing personnel costs (compared to the usual staffing levels), which are acceptable given the large reductions in understaffing risks.

In the case of *average* ED census forecasts, once again, M5P was the best choice for the reduction of major and medium underestimation risks, for all forecast horizons. Relative risk reductions were similar to those of *maximum* ED census forecasts (more than 10-fold reduction in major underestimation and approximately 3-fold reduction in medium underestimation, compared to *usual resource allocation policies*). However, in this case, the usual resource allocation policies already had low risks of major and medium underestimation (~2% risk of major underestimation and ~3.7% risk of medium underestimation). On the other hand, most importantly, in the case of average ED census forecasts, M5P led to a *cost reduction* of more than 15% compared to the usual resource allocation policies.

The second topic of this Ph.D. thesis is the development of models for real-time prediction of probabilities of inpatient admission from the ED. Our aim in this case was the development of classifiers with adequate performance in terms of both discrimination and calibration (goodness-of-fit), reliant on a small number of variables, available in most ED settings right

after triage. In our setting, the Manchester Triage System (MTS) was used. Discrimination was evaluated with the area under the ROC curve (AUROC). Calibration was evaluated with Hosmer-Lemeshow (H-L) χ^2 and p-values with 10 fixed probability intervals. We used logistic regression (LR) models, artificial neural networks (ANN) models and models based on an ad hoc ensemble classifier that optimized calibration (it combined a LR model with a tree of MTS chief complaints with LogitBoost on its leaves). A custom method was used for the evaluation of models, with increasingly larger train sets and 12 consecutive test sets of approximately monthly length. This evaluation method produced the results that follow, reported with 95 % confidence intervals (CIs). For LR models, average AUROC = 0.8531, 95% CI (0.8501, 0.8561); for ANN models, average AUROC = 0.8568, 95% CI (0.8531, 0.8606) and for ad hoc ensemble classifier models, average AUROC = 0.8635, 95% CI (0.8605, 0.8665). Confidence intervals of average AUROCs for LR and ad hoc ensemble classifier models did not overlap. Confidence intervals of average AUROCs for LR and ANN models slightly overlapped, although ANN models had higher AUROCs than LR models in all but one of the 12 test sets. Average H-L χ^2 were, respectively, 35.15, 95% CI (32.57, 37.73) for LR models, 10.47, 95% CI (7.78, 13.17) for ANN models and 11.4, 95% CI (9.10, 13.75) for ad hoc ensemble classifier models. Both ANN and ad hoc ensemble classifier models possessed better calibration than LR models, with H-L p-values > 0.05 in 10 of the 12 experiments.

The third topic of this Ph.D thesis is the development and evaluation of software for the generation of logistic and Cox regression nomograms. We developed two programs (*nomolog* and *nomocox*) for these purposes, based on Stata (a statistical software package widely used in biomedical research). At the time of the writing of this Ph.D. thesis these programs are used by an international community of researchers in the fields of clinical medicine, epidemiology or biostatistics. We surveyed some of these users about their background, their user experience with *nomolog* and *nomocox*, as well as the ease-of-use and flexibility of our programs compared to those available in R (another well-known statistical software). Most respondents were “Promoters” (Net Promoter Score > 8), i.e. very likely to recommend the software to other researchers. All respondents (100%) who had used both our programs (*nomolog* and *nomocox*) and nomogram generators available for the R statistical software, found *nomocox* and *nomolog* easier to use; with a 95% adjusted Wald CI (75.83%, 100%). A raw proportion of 81.25%, with a 95% adjusted Wald CI (54.03%, 96.36%) found our programs to be more flexible than the nomogram generators available for the R statistical software.

Resumen

El Sistema Nacional de Salud (SNS) español es fundamentalmente público, tanto en su financiación, como en la titularidad de los entes proveedores. El SNS español es considerado altamente eficiente en términos de costes según estudios internacionales basados en datos de la Organización Mundial de la Salud (OMS). Sin embargo, la contención del incremento de los costes sanitarios, manteniendo a la vez niveles de calidad asistencial adecuados, sigue siendo un problema de difícil solución. En los últimos años los servicios de urgencias de los hospitales de atención especializada se han visto sometidos a restricciones presupuestarias, acompañadas de un incremento simultáneo del número de visitas y de la complejidad clínica de las mismas. Estas circunstancias exigen nuevos modelos de gestión de estos servicios de urgencias, que podrían beneficiarse de sistemas de soporte a la decisión.

En esta tesis doctoral se proponen soluciones basadas en algoritmos de aprendizaje automático para dos problemas comunes a la mayoría de los servicios de urgencias de hospitales de atención especializada: la predicción del censo de urgencias y la predicción en tiempo real de las probabilidades de hospitalización para todos los pacientes triados presentes en el servicio de urgencias en un determinado momento. Dichas soluciones pueden emplearse en la construcción de sistemas de soporte a la decisión. Los datos para dichos estudios fueron proporcionados por el Hospital Universitario Ramón y Cajal de Madrid, un hospital de atención especializada con todas las especialidades médicas exceptuando Obstetricia. En los años 2011 y 2012 este hospital tenía aproximadamente 1.100 camas instaladas y aproximadamente 553.000 pacientes asignados a su área clínica. Otro tema tratado en esta tesis es el desarrollo y la evaluación de herramientas software para la generación de nomogramas para modelos basados en regresión logística y en regresión de Cox, dado que los nomogramas pueden emplearse para ayudar en la toma de decisiones clínicas o en procedimientos de contingencia en caso de fallo de sistemas de soporte a la decisión basados en sistemas de información.

El primer tema de esta tesis doctoral es la generación de modelos para la predicción de niveles de censo de urgencias (número de pacientes en el servicio de urgencias en un determinado instante de tiempo). Una de las aplicaciones de la predicción del censo de urgencias es la planificación del personal de enfermería, basada en recomendaciones nacionales e internacionales. Se decidió emplear una granularidad de 8 horas en las predicciones dado que numerosos recursos de los servicios de urgencias hospitalarias están organizados en turnos de 8 horas. Nuestro objetivo era generar predicciones para dos

variables dependientes: niveles máximos de censo de urgencias y niveles medios de censo de urgencias. Las predicciones de niveles máximos de censo de urgencias en turnos de 8 horas pueden emplearse para la planificación de personal de enfermería, mientras que las predicciones de niveles medios de censo de urgencias en turnos de 8 horas pueden emplearse para otros fines (tales como la asignación de personal administrativo).

Se construyeron modelos de series temporales basados en métodos de regresión generalizada con los siguientes algoritmos de aprendizaje automático: M5P, Alternating Model Trees (AMT) y Support Vector Regression (SVR). Estos modelos fueron comparados con una serie de modelos de referencia: dotación de personal habitual (y niveles de dotación de recursos habituales), medias estratificadas (medias estratificadas por los tres turnos diarios, de 8 horas cada uno), regresión lineal y Seasonal Autoregressive Integrated Moving Average (SARIMA). Se generaron predicciones para las dos variables dependientes: niveles medios de censo de urgencias y niveles máximos de censo de urgencias. Se probaron cuatro horizontes de predicción: 1 semana, 2 semanas, 4 semanas y 8 semanas. Se definieron riesgos de subestimación, riesgos de sobreestimación y aproximaciones a costes de políticas de asignación de recursos para las predicciones de niveles de censo medios y máximos. Las predicciones de niveles de censo máximos fueron transformadas en requisitos de personal de enfermería; y los riesgos de subestimación y sobreestimación para niveles de censo máximos fueron transformados en riesgos de subestimación y sobreestimación de personal. Se empleó un esquema propio de entrenamiento y prueba, con conjuntos de entrenamiento de tamaño creciente y conjuntos de prueba de tamaño fijo. También se empleó el mismo esquema con una variación consistente en el uso de conjuntos de entrenamiento de tamaño fijo de un año y conjuntos de prueba de longitud fija. Este último esquema no produjo una mejora de los resultados.

En el caso de las predicciones de niveles máximos de censo de urgencias, M5P fue la mejor opción para la reducción del riesgo de situaciones de subestimación de severidad alta y severidad media de personal de enfermería para todos los horizontes de predicción. En comparación con los niveles habituales de personal de enfermería, el uso de M5P podría reducir los riesgos de severidad alta de subestimación de personal de enfermería (> 9 enfermeros/as) en más de 10 veces (una reducción hasta el ~1% con M5P en comparación con el ~13% obtenido con los niveles habituales de personal de enfermería); y podría reducir los riesgos de subestimación de severidad media de personal de enfermería (7-9 enfermeros/as) en aproximadamente 3 veces (una reducción hasta el ~3 % con M5P en comparación con el ~10% obtenido con los niveles habituales de personal de enfermería). El uso de M5P también implicaría incrementos de costes estimados de personal de enfermería

de 5% - 6,1% en comparación con los niveles de personal habituales. Dicho incremento de costes sería aceptable dadas las grandes reducciones de riesgos de subestimación de personal de enfermería.

En el caso de las predicciones de niveles medios de censo de urgencias, M5P también fue la mejor opción para la reducción de los riesgos de subestimación de severidad alta y media para todos los horizontes de predicción. Las reducciones relativas de riesgos de subestimación eran similares a las de las obtenidas sobre las predicciones de niveles máximos de censo de urgencias (reducción en más de 10 veces de los riesgos de subestimación de severidad alta y reducción en aproximadamente 3 veces de los riesgos de subestimación de severidad media, en comparación con las políticas habituales de asignación de recursos). Sin embargo, en este caso, las políticas habituales de asignación de recursos ya presentaban riesgos reducidos de subestimación de severidades alta y media ($\sim 2\%$ de riesgo de subestimación de severidad alta y $\sim 3,7\%$ de riesgo de subestimación de severidad media). No obstante, en el caso de las predicciones de niveles medios de censo de urgencias, el uso de M5P conduciría también a una *reducción de costes* superior al 15% en comparación con las políticas habituales de asignación de recursos.

El segundo tema tratado en esta tesis doctoral es el desarrollo de modelos para la predicción de probabilidades de hospitalización procedente de urgencias en tiempo real. El objetivo en este caso era desarrollar modelos predictivos basados en un número reducido de variables disponibles en la mayor parte de los servicios de urgencias de atención especializada inmediatamente después del triaje. En nuestro caso se empleaba el Sistema de Triage Manchester (MTS). Asimismo, era fundamental diseñar clasificadores con un rendimiento adecuado en términos de discriminación y calibración. La discriminación se evaluó con áreas bajo la curva ROC (AUROC). La calibración se evaluó con χ^2 y p-valores de Hosmer-Lemeshow (H-L), con 10 grupos fijos de probabilidad. Se emplearon modelos de regresión logística (RL), redes neuronales artificiales (RNA) y modelos basados en un algoritmo ad hoc que optimizaba la calibración (combinaba un modelo base de regresión logística con un árbol de motivos de consulta del MTS con modelos LogitBoost en sus ramas). Se empleó un método de evaluación propio, con conjuntos de entrenamiento de tamaño creciente y 12 conjuntos de prueba consecutivos de una longitud aproximadamente mensual. Dicho método de evaluación produjo los resultados que se indican a continuación, con intervalos de confianza (ICs) al 95 %. Para los modelos de RL, AUROC media = 0,8531, IC 95% (0,8501 , 0,8561), para los modelos de RNA, AUROC media = 0,8568, IC 95% (0,8531 , 0,8606), y para los modelos ad hoc, AUROC media = 0,8635, IC 95% (0,8605 , 0,8665). Los intervalos de confianza de AUROCs medias para modelos de RL y ad hoc no se superponían. Los

intervalos de confianza de AUROCs medias para modelos de RL y modelos de RNA sí se superponían ligeramente, aunque los modelos de RNA tenían AUROCs más elevadas que los modelos de RL en los 12 conjuntos de prueba exceptuando uno. Las χ^2 de H-L promedio fueron, respectivamente, 35,15 IC 95% (32,57 , 37,73) para modelos de LR, 10,47; IC 95% (7,78 , 13,17) para los modelos de RNA y 11,4, IC 95% (9,10 , 13,75) para los modelos ad hoc. Tanto los modelos de RNA como los modelos ad hoc tenían una calibración mejor que los RL, con p-valor H-L > 0.05 en 10 de los 12 experimentos.

El tercer tema de esta tesis doctoral es el desarrollo y evaluación de software para la generación de nomogramas basados en modelos de regresión logística binaria y regresión de Cox. Se han desarrollado dos programas (*nomolog* y *nomocox*) basados en Stata (un paquete de software estadístico ampliamente usado en la investigación biomédica). En el momento de la redacción de esta tesis doctoral, estos programas son usados por una comunidad internacional formada por investigadores en las disciplinas de medicina clínica, epidemiología y bioestadística. Se realizó una encuesta a dicha comunidad de usuarios sobre el perfil de cada investigador, así como sobre la experiencia de usuario de los programas *nomolog* y *nomocox*. También se evaluó la facilidad de uso y la flexibilidad de estos programas en comparación con los generadores de nomogramas disponibles en el software estadístico R (un software estadístico ampliamente conocido). La mayoría de los encuestados afirmaron ser "promotores" (Net Promoter Score > 8), es decir, muy propensos a recomendar el software a otros investigadores. Todos los encuestados (100%), que habían utilizado tanto los generadores de nomogramas disponibles para el software estadístico R como los generadores de nomogramas *nomolog* y *nomocox*, indicaron que estos últimos eran más fáciles de usar, con un IC 95% de Wald ajustado (75,83% , 100%). Asimismo, una proporción del 81,25%, IC 95% de Wald ajustado (54,03% , 96,36%) afirmó que *nomolog* y *nomocox* eran más flexibles que los generadores de nomogramas disponibles para el software estadístico R.

Table of Contents

Institutional acknowledgments	7
Abstract	9
Resumen	12
Index of figures	19
1. Introduction	23
1.1. Background and motivation	23
1.2. Objectives	27
1.2.1. General objectives	27
1.2.2. Specific objectives	28
1.2.3. Impact and applicability	31
1.3. Thesis structure	32
2. Emergency department (ED) census forecasting	34
2.1. Forecasts of ED visits and census levels	34
2.1.1. Introduction	34
2.2. State of the art	35
2.2.1. Previous approaches	35
2.2.2. Autoregressive integrated moving average and seasonal integrated autoregressive moving average models	37
2.2.3. Machine learning time series models	39
2.2.4. Risks measurements in forecasting problems for ED nursing personnel allocation	44
2.3. Materials and Methods	45
2.3.1. Data source	45
2.3.2. Dependent variables: ED visits <i>versus</i> ED census levels	46
2.3.3. Independent variables	50
2.3.4. Performance measurements: underestimation risks, overestimation risks and costs	52
2.3.5. Forecasting algorithm parameterization and custom time series evaluation scheme	55
2.4. Results	76
2.4.1. Maximum ED census levels	76
2.4.2. Average ED census levels	82
2.4.3. Evaluation with fixed-length train window	84
2.5. Discussion	88
2.5.1. Limitations	91
2.6. Conclusions	92
3. Prediction of inpatient admission from the emergency department (ED)	94
3.1. Introduction	94

3.1.1.	The case for decision support systems in Medicine.....	94
3.1.2.	Inpatient admissions from the ED and ED triage systems	96
3.1.3.	The Manchester Triage System (MTS)	98
3.1.4.	The prediction of inpatient admission immediately after triage.....	100
3.2.	State of the art.....	104
3.2.1.	Calibration versus discrimination. Hosmer-Lemeshow “groups of risk”	104
3.2.2.	Algorithms capable of producing adequately calibrated models.....	113
3.2.3.	Previous research.....	123
3.3.	Materials and Methods.....	126
3.3.1.	Data source and study setting.....	126
3.3.2.	Optimization objective	128
3.3.3.	Independent variables included <i>a priori</i>	130
3.3.4.	Independent variables excluded <i>a priori</i>	135
3.3.5.	Feature space exploration.....	138
3.3.6.	ANN and LogitBoost parameterization.....	142
3.3.7.	Ad hoc classifier derivation	149
3.4.	Results	154
3.4.1.	Logistic regression odds ratios	154
3.4.2.	Calibration and discrimination on the test set.....	156
3.4.3.	Estimations of calibration and discrimination metrics with increasing train sets and fixed-length test sets.....	160
3.5.	Discussion.....	164
3.5.1.	Interpretation of results	164
3.5.2.	Possibilities of deployment of predictive system	166
3.5.3.	Limitations	167
3.6.	Conclusions	168
4.	Nomograms	169
4.1.	Introduction.....	169
4.1.1.	Nomograms	169
4.1.2.	Logistic regression nomogram example	173
4.1.3.	Derivation of logistic regression nomograms	175
4.1.4.	Derivation of Cox regression nomograms.....	181
4.1.5.	Example of a nomogram for clinical practice: evolution of back pain	183
4.2.	User survey	184
4.2.1.	Introduction.....	184
4.2.2.	Materials and methods.....	185
4.2.3.	Results	193
4.2.4.	Discussion.....	199
4.3.	Conclusions	200
5.	Conclusions, Contributions and Future work	202

5.1. Conclusions	202
5.1.1. Emergency department census forecasting.....	202
5.1.2. Prediction of inpatient admission from the emergency department.....	203
5.1.3. Logistic and Cox regression nomograms.....	206
5.2. Contributions	207
5.2.1. Emergency department census forecasting.....	207
5.2.2. Prediction of inpatient admission from the emergency department.....	209
5.2.3. Logistic and Cox regression nomograms.....	211
5.3. Future work	212
5.3.1. Emergency department census forecasting.....	212
5.3.2. Prediction of inpatient admission from the emergency department.....	214
5.3.3. Logistic and Cox regression nomograms.....	218
6. Publications	219
6.1. International scientific journals (primary)	219
6.2. International scientific journals (secondary)	219
6.3. International scientific conferences (primary)	220
6.4. International scientific conferences (secondary)	221
References	222
Appendix A. Average ED census forecasts compared to personnel needs indicated by maximum ED census forecasts.....	233
Appendix B. Error distributions for inpatient admission prediction for the ad hoc ensemble classifier	238
Appendix C. Nomogram usage	240
Appendix D. Software developed for this Ph.D. thesis	247

Index of figures

Figure 1. Emergency department visits and Emergency department census.	47
Figure 2. Maximum and average hourly census levels in 8-hour shifts.	48
Figure 3. Time series evaluation scheme.	58
Figure 4. Average of MAEs for "average ED census" forecasts with the "basic" group of exogenous variables and varying <i>maximum lags</i>	62
Figure 5. Average of MAEs for "maximum ED census" with the "basic" group of exogenous variables and varying <i>maximum lags</i>	63
Figure 6. Average of MAEs for "average ED census" with the "basic" group of exogenous variables, maximum lags = 1 and varying <i>parameter M</i>	65
Figure 7. Average of MAEs for "maximum ED census" with the "basic" group of exogenous variables, maximum lags = 1 and varying <i>parameter M</i>	66
Figure 8. MAE on the validation set for the prediction of average ED census with a forecast horizon of 84 shifts (4 weeks ahead) for an AMT algorithm with variable parameters H and I ($H=0..1$; $I=0..160$). The marker ($X=Parameter\ I\approx 5$; $Y=parameter\ H=1$; $Z=MAE\approx 8.4$) is set at an optimum which is deemed unstable. The red zone is assumed to be a stable optimum.	69
Figure 9. MAE on the (A) train set and (B) validation set as a function of parameters C and E of an SVR time series for the prediction of <i>maximum census</i> and a forecast horizon of 84 shifts (4 weeks ahead).	71
Figure 10. MAE for a forecast of "average ED census" for an SVR algorithm with fixed parameter $E=1.0$ (linear kernel) and variable parameter C (regularization).	72
Figure 11. MAE for a forecast of maximum ED census for an SVR algorithm with fixed parameter $E=1.0$ (linear kernel) and variable parameter C (regularization).	73
Figure 12. Estimated risks of understaffing (calculated upon <i>maximum</i> ED census forecasts) for different algorithms and forecast horizons.	79
Figure 13. Estimated risks of overstaffing (calculated upon <i>maximum</i> ED census forecasts) for different algorithms and forecast horizons.	80
Figure 14. Estimated yearly nursing personnel costs (calculated upon <i>maximum</i> ED census forecasts) for different algorithms and forecast horizons.	81
Figure 15. Estimated risks of underestimation (calculated upon <i>average</i> ED census forecasts) for different algorithms and forecast horizons.	86
Figure 16. Estimated risks of overestimation (calculated upon <i>average</i> ED census forecasts) for different algorithms and forecast horizons.	87
Figure 17. Inpatient admission prediction <i>immediately after triage</i>	101
Figure 18. ROC curve of a binary classifier with only two output values.	105
Figure 19. Examples of calibration graphs. (A) Circles of equal size; (B) Circle sizes used to represent N_k	111
Figure 20. McCulloch and Pitts neuron.	119
Figure 21. Perceptron.	120

Figure 22. Comparison of model performance on the (A) train and (B) validation datasets. <i>108 models with less than 8 degrees of freedom are excluded on the train dataset (A) and 135 models are excluded on the validation dataset (B).</i>	141
Figure 23. Comparison of <i>patternnet</i> ANN H-L χ^2 on the (A) train and (B) validation datasets. 0-50 hidden neurons and 0-1 regularization. <i>Only models with 8 degrees of freedom are shown.</i>	145
Figure 24. Comparison of <i>patternnet</i> ANN AUROC on the (A) train and (B) validation datasets. 0-50 hidden neurons and 0-1 regularization.....	146
Figure 25. Evolution of LogitBoost performance on train and validation datasets (A) AUROC; (B) H-L χ^2 from 0 to 400 iterations; (C) H-L χ^2 from 50 and 200 iterations.	148
Figure 26. Calibration on the train set. (A) Logistic regression; (B) ANN; (C) Ensemble classifier ..	158
Figure 27. Nomogram representation for the logistic regression model.	159
Figure 28. Logistic regression and Ensemble classifier evaluation scheme.	162
Figure 29. Comparison of Logistic regression, ANN and Ensemble classifier on an increasingly larger train set, with a 1 month test set in all cases. (A) AUROC; (B) H-L χ^2 with 8 degrees of freedom. ...	163
Figure 30. Logistic regression nomogram example.....	174
Figure 31. Nomogram for low back pain clinically relevant improvement.	183
Figure 32. Questionnaire response processing scheme.	193
Figure 33. Answers to Q.19, Q.20 and Q.21.	198

Index of tables

Table 1. Definition of bank holiday weekend independent variables.	51
Table 2. Overstaffing and understaffing scenarios for <i>maximum</i> ED census forecasts.	52
Table 3. Usual resource allocation policies in the ED of the Ramón y Cajal University Hospital.	53
Table 4. Overestimation and underestimation errors defined for <i>average</i> ED census forecasts.	54
Table 5. Parameterization datasets	56
Table 6. ED census levels experiments: independent variables and forecast horizons.	59
Table 7. M5P parametrization strategy	60
Table 8. Optimum M5P parametrization on the validation set.	60
Table 9. AMR and SVR parametrization strategy.	67
Table 10. AMT initial parametrization.....	68
Table 11. AMT reviewed parametrization. <i>Differences with initial parametrization in bold.</i>	69
Table 12. Results on the macro test set (4 years) with algorithms for <i>maximum</i> ED census. <i>95% CI overlapping with those of M5P are highlighted in bold.</i>	78
Table 13. Results on the macro test set (4 years) with algorithms for <i>average</i> ED census. <i>95% CIs overlapping with those of M5P highlighted in bold.</i>	85
Table 14. Manchester triage system's priority levels.	99
Table 15. The effect of sample size on H-L calibration tables and H-L χ^2 values based on <i>fixed</i> probability intervals. (A) Large sample size; (B) Medium sample size; (C) Small sample size.	110
Table 16. The effect of the number of groups (g) on H-L calibration tables and H-L χ^2 values based on fixed probability intervals.....	112
Table 17. Similar previous models for the prediction of inpatient admission from the Emergency Department.....	125
Table 18. Classification of chief complaints by risk of admission on the train set.	132
Table 19. Odds ratios for the LR model built on the train and validation sets.	155
Table 20. Questionnaire for users of the <i>nomolog</i> and <i>nomocox</i> programs.	187
Table 21. Affiliations of the researchers who answered the questionnaire.....	194
Table 22. Demographics, researcher profile, knowledge of logistic & Cox regression, knowledge of Stata, R and general statistical software.	195
Table 23. Usage given to logistic and Cox regression nomograms by the respondents. <i>LCI = Lower confidence interval; UCI = Upper confidence interval.</i>	196

Glossary of Terms and Abbreviations

ED = Emergency Department (of a specialized care hospital)
(inpatient) admissions = Patients transferred to hospitalization wards
HIS = (central) Hospital Information System
NHS = National Healthcare System
MTS = Manchester Triage System
CPU = Central Processing Unit
GPU = Graphical Processing Unit
R = R statistical software
Stata = Stata statistical software
SAS = Statistical Analysis Software
ARIMA = Autoregressive integrated moving-average
SARIMA = Seasonal ARIMA
ARCH = Autoregressive conditional heteroskedasticity
GARCH = Generalized ARCH
MAE = Mean absolute error
MSE = Mean squared error
RMSE = Root mean squared error
TPR = True positive rate
FPR = False positive rate
H-L “deciles of risk” = Hosmer-Lemeshow “deciles of risk”
H-L test = Hosmer-Lemeshow test of calibration
AUROC = Area under the receiver operating characteristic (ROC) curve
CI = Confidence interval
UCI = Upper confidence interval
LCI = Lower Confidence interval
CART = Classification and regression tree
C4.5 = A classification tree algorithm
M5P = A regression tree based algorithm
LR = Logistic regression (binary)
ANN = Artificial neural network
SVM = Support vector machine
SVR = Support vector (machine) regression

1. Introduction

1.1. Background and motivation

The right to healthcare was defined in the Spanish constitution of the year 1978 and a central authority - the Instituto Nacional de la Salud (INSALUD) - was created to manage all public healthcare institutions. The General Healthcare Act of 1986 defined a system of near-universal healthcare coverage following the Beveridge model and set the roadmap for the transfer of healthcare responsibilities to the Spanish Regions (Autonomous Communities). This process was completed in 2001. From then on, the functions of the Ministry of Healthcare and Social Services (MoHSS) were reduced to such activities as healthcare policy coordination, statistical analyses and inter-regional retribution through a “Cohesion Fund” [1, 2].

At present, the Spanish National Healthcare System (NHS) is mostly publicly funded and publicly provided. It is considered to be highly cost-efficient according to several studies based on World Health Organization data [3-5]. However, as in most developed countries, increasing healthcare costs have remained a significant concern [6]. During the Spanish financial crisis that began in 2008, new laws introduced restrictions to healthcare coverage and increased co-payments [7]. In most Autonomous Communities, healthcare budgets were contained or reduced [8], but the large differences in healthcare expenditures per inhabitant between Autonomous Communities remained in place [8].

Primary care centers and specialized care hospitals constitute the backbone of the Spanish NHS. Emergency Departments (EDs) of specialized care hospitals are one of the areas of the healthcare system which has experienced most pressure in recent years due to budget

restrictions combined with an increase in patient visits and clinical complexity. There has been an increase in patient presentations due to (i) a significant reduction in emergency care services in primary care centers [9], (ii) increased waiting times for primary care consultations [10], (iii) restrictions in access to primary care for persons without Social Security insurance (healthcare coverage has been reduced since 2012) [11]. At the time of the writing, visits to specialized care EDs are more frequent in Spain than in many other European countries [12]. For example, there were approximately 436 visits per 1000 inhabitants in 2013 for the entire Spanish healthcare system [13]. At the same time, patient complexity has increased due to population aging [14], among other causes.

An increased demand with a similar level of resources is likely to cause overcrowding in any system, and EDs are no exception. Causes of ED overcrowding are usually classified in input factors (sources and aspects of patient visit flow), throughput factors (bottlenecks within the ED) and output factors (bottlenecks in other parts of the healthcare system that might affect the ED) [15]. The consequences of ED overcrowding are: excessive waiting times, increased patient complaints, patients leaving without being seen [16], boarded patients (bed-ridden patients waiting in the hallways), ambulance diversion, longer hospital length of stay [17], ethical dilemmas [18], staff burnout as well as increased clinical errors, which may cause negative clinical outcomes [19].

The main problems of ED management are the variability of demand [20] and the high risk of inadequate resource provision. However, ED management is also a promising application domain for healthcare operations research. Particularly, models that can be used as decision support systems can be helpful for several managerial problems. For instance, time-series models can be developed to forecast ED census levels with sufficient advance as to allow for ED resource planning. Also, a system capable of accurately establishing the probabilities of inpatient admission (hospitalization) for every ED patient right after triage

can help streamline operations and establish priorities for clinical personnel, bed managers and supporting personnel. Previous research had shown that clinical *triage*¹ personnel could not predict the need for inpatient admission with sufficient reliability [21-23], however, models with a manageable number of easily-obtainable variables and a simple procedure for calculating the probabilities of admission could be used to aid in this task [24-26]. In both cases (ED census forecasts and prediction of probabilities of inpatient admission from the ED) predictive models with varying degrees of sophistication can be developed.

Although numerous forecasting and supervised prediction techniques have been proposed for the aforementioned problems, to the best of our knowledge, several relevant issues have not been addressed in previous research:

(i) In time series models developed for ED personnel planning, the granularity of predictions was often insufficient. For instance, daily visit numbers are of little use if clinical personnel staffing is usually organized in 8-hour shifts, however, almost all previously developed models used a daily (24-hour) granularity for the forecasts. Also, forecast horizons were often insufficient to adapt staffing levels to forecasts.

(ii) Previous models developed for the prediction of inpatient admissions from the ED have often relied on variables not available in all EDs of specialized care hospitals at all or with sufficient advance. Examples of these are the total number of pharmacy dispensations in primary care [24] for a given patient, or standardized chronic clinical conditions [25], which are not available in all EDs of specialized care hospitals in real time (or at all). Also, several models relied on variables which were highly specific to a country or a region [25, 26], which made them impossible to apply in other regional settings.

¹ Triage: initial clinical assessment performed in the ED immediately after administrative check-in.

(iii) Predictive models developed in the course of clinical research are often presented in such a way that it is very laborious (or impossible) to obtain the output variable given a set of input variable values. A typical example is a table of odd-ratios of variables (but not the constant term) for a predictive model based on a logistic regression ². This common representation provides no way of calculating the probability of event. On the other hand, the full regression formula, which is sometimes provided in appendices [25], most often requires laborious calculations. Nomograms are representations of logistic and Cox regression models that allow easy calculations of output probabilities. Logistic and Cox regression nomograms can also be used as backup procedures for decision support systems based on complex machine learning algorithms. However, until now, there was no user-friendly software for logistic and Cox regression nomogram generation.

² Unless otherwise specified, the term “logistic regression” is used to refer to a “*binary* logistic regression” in this Ph.D. thesis.

1.2. Objectives

1.2.1. General objectives

The objective of this thesis was to develop tools that could be used to build decision aids and decision support systems for applications related to Emergency Department (ED) management. We intended to build predictive models based on machine learning classification techniques and machine learning time series techniques. Since our aim was to maximize applicability and portability, we sought to rely only on independent variables available in most specialized care ED settings. We also wanted to emphasize the usage of realistic error metrics and performance evaluation methods for the predictive models devised in this work. We also aimed to build software tools for the generation of logistic and Cox regression nomograms, since these can be useful in clinical decision making, research and teaching.

The data for the development of classification and forecasting models were provided by a large specialized care referral center of the Madrid Region - the Ramón y Cajal University Hospital.

The main sub-objectives of the thesis are stated in the subsequent pages.

1.2.2. Specific objectives

1.2.2.1. Design, implementation and evaluation of autoregressive and machine learning time series models for emergency department census forecasting

Our primary objective was the creation and comparison of predictive time series models for nursing staff allocation in Emergency Departments (EDs) of specialized care hospitals. Our aim was to reduce major and medium understaffing situations of nursing personnel with a cost increase of less than 10% (compared to the usual resource allocation policies). As international nurse staffing ratios for a predefined number of patients exist [20, 27-32], forecasts of patients present at the ED at a given moment (ED census levels) could be transformed into nursing personnel requirements. As a secondary objective, we also sought to develop predictive time series models which could serve for the allocation of resources based on average ED census levels. As with maximum ED census forecasts, our goal was to reduce major and medium underestimation situations with cost increases of less than 10%, compared to the usual resource allocation policies. We chose an 8-hour granularity for our forecasts since many resources in the EDs were organized in 8-hour shifts.

Generalized regression time series models based on machine learning algorithms were to be developed and to be compared to a series of benchmarks: usual resource allocation policies, stratified averages, linear regression time series and Seasonal Autoregressive Integrated Moving-Average (SARIMA) models. The performances of the different models were to be evaluated with mean absolute error (MAE) as well as risks of overestimation and risks of underestimation. Approximate monetary costs of resource allocation policies, derived from the forecasts, would be also included in the evaluation of results.

1.2.2.2. Design, implementation and evaluation of machine learning classification models for the prediction of inpatient admission from the emergency department

We developed models for the real-time prediction of probabilities of hospitalization from the ED (also called *inpatient admission* from the ED) for all triaged patients, excepting those with chief complaint “Pregnancy” (since the Ramón y Cajal University Hospital had all medical specialties excepting Obstetrics). Such systems could improve clinical and administrative patient flow by giving the possibility of allocating resources in advance. They could also be helpful for tasks often performed or aided by nursing and medical personnel such as fast-track admission, inpatient and ED bed management, rapid discharge, and outbound transfer planning. The predictions produced by these systems could also benefit administrative and supporting personnel in the ED and could be integrated into existing ED workflows. The underlying models were to be based on variables available immediately after triage in most ED settings.

Machine learning classification algorithms (ANNs, LogitBoost and classification trees with logistic regression models on its leaves) would be used for the development of the predictive models. We also intended to design a custom ad hoc classifier (an ensemble of a logistic regression model and a tree of triage chief complaints with LogitBoost on its leaves) in order to further improve our results. We also sought to develop an adequately calibrated logistic regression model with a nomogram representation. This nomogram could then serve as a paper-based contingency procedure in case of failure of software decision support systems.

Our aim was to produce models with an AUROC > 0.85 , decreasing H-L χ^2 as much as possible (with 10 fixed probability groups), reaching H-L p-value > 0.05 if possible.

1.2.2.3. Design, implementation and evaluation of nomogram generators for logistic and Cox regression models

Logistic and Cox regression nomograms are graphical representations of predictive models based on logistic and Cox regressions, that (i) can serve as aids for clinical decision-making; (ii) may be used in contingency methods in case of failure of software decision support systems; (iii) can improve the usability of predictive models included in scientific publications; (iv) may be used for teaching purposes.

To the best of our knowledge, when we started our work on this Ph.D. thesis, there was no user-friendly and flexible software available for the generation of nomograms based on logistic and Cox regressions. Our aim was develop such software. It would be based on Stata – a statistical software widely used by biomedical researchers and clinicians. We intended to develop nomogram generators based on Stata for both logistic and Cox regression models, since these were arguably the most common regression techniques in biomedical research at the time of the writing. We also intended to survey the users of the software we aimed to develop in order to evaluate user satisfaction and to compare our software with existing nomogram generators available for the R statistical software.

1.2.3. Impact and applicability

Our aim was to develop predictive models (for ED census forecasting and for the prediction of probabilities of inpatient admission from the ED) that could be used in most EDs of specialized care hospitals in Spain, although adjustments could be necessary. We intended to develop models that could also be used in specialized care hospitals in other countries, since it was deemed possible to obtain variables equivalent to the ones used in our models in most ED settings in other countries.

We believe that the usage of ED census forecasting models presented in this thesis can result in reduced risks of underestimation of resources compared to classical estimation methodologies (such as rule-of-thumb methods, stratified averages or SARIMA models). We also believe that it may reduce costs of resource allocation policies in some settings.

The models developed for the real-time estimation of probabilities of inpatient admission (hospitalization) from the Emergency Department can serve to manage patient flow in a more efficient manner and as an early warning system for clinical personnel in the ED and in the hospitalization wards.

The nomogram generators for logistic and Cox regressions can be of use for clinical researchers. Such nomograms are commonly used for clinical decision-making, communication of scientific results and teaching purposes. At the time of the writing, the programs that we have developed are used in several international clinical research institutions such as the Karolinska Institute, the Harvard Medical School or the Chinese Academy of Medical Sciences.

1.3. Thesis structure

The structure of this thesis is as follows:

- Chapter 1 includes the background and motivation, a description of the objectives of the thesis, as well as the impact and applicability of this work.
- Chapter 2 describes the forecasting of ED census levels stratified by 8-hour shifts with machine learning times series models. Forecasting performance is benchmarked against *usual resource allocation policies*, stratified averages, linear regression time series and SARIMA. The data for the forecasting models were provided by a large specialized care referral center of the Madrid Region - the Ramón y Cajal University Hospital.
- Chapter 3 describes the prediction of inpatient admissions (hospitalizations) from the ED by means of logistic regression, ANN and an ad hoc ensemble classifier (composed by a logistic regression model and LogitBoost submodels for triage chief complaints). The data for the development of the classification models were also provided by the Ramón y Cajal University Hospital.
- Chapter 4 describes the design of two software programs for the generation of logistic and Cox regression nomograms (*nomolog* and *nomocox*) and a survey of users of these programs.
- Chapter 5 includes a summary of conclusions, contributions and a description of proposed future research related to the topics addressed in this Ph.D. thesis.

- Chapter 6 lists the scientific publications completed in the course of this Ph.D. thesis: articles in national scientific journals, articles in international scientific journals, and presentations in national and international scientific conferences.
- Appendix A contains an analysis of errors, risks and costs if *average* ED census forecasts were used for nursing personnel allocation when the actually needed number should have been established according to *maximum* ED census.
- Appendix B includes graphical representations of error distributions on a large test set for the ad hoc ensemble classifier developed for the prediction of probabilities of inpatient admission from the ED.
- Appendix C discusses several examples of usage of logistic and Cox regression nomograms.
- Appendix D lists the software packages developed for this Ph.D. thesis.

2. Emergency department (ED) census forecasting

2.1. Forecasts of ED visits and census levels

2.1.1. Introduction

Reliable forecasts of ED census levels within 8-hour shifts could be of great use for ED staff planning. Recommendations for nurse staffing requirements depending on the expected number of patients in the ED exist in Spain [20], although the application of these standards is not legislatively enforced, as, for instance, in the state of California (United States of America) [27]. In Spain, hospital management teams most often employ rule-of-thumb estimations for nursing personnel needs [20], which may deviate from these standards due to the lack of reliable predictions of ED census levels (number of patients in the ED at a given moment). We believe that the usage of predictive models based on machine learning approaches for nurse staff allocation (based on *maximum* ED census levels within 8-hour shifts) is possible and could significantly reduce risks of major and medium understaffing with cost increases of less than 10% (compared to the usual resource allocation policies). Our primary objective was to develop such models.

As a secondary objective, we also intended to develop models that could serve for the allocation of resources in the ED based on *average* ED census levels within 8-hour shifts. For these, our optimization objective was also a significant reduction of risks of major and medium underestimation with cost increases of less than 10%.

2.2. State of the art

2.2.1. Previous approaches

Accurately predicting ED patient flow with sufficient advance and granularity can be challenging. Real-time crowding assessment tools have been devised [33-35], but their practical usefulness has been contested [36]. Short-term (up to 12 hours ahead) crowding forecasting models have also been developed [37-39]. These usually rely on such factors as patient numbers stratified by triage score and numbers of attending physicians at a given moment in the ED. Some authors have also employed variables such as the number of radiological and laboratory tests in the preceding hours [38]. Also, short-term forecasting Seasonal Autoregressive Integrated Moving Average (SARIMA) models for inpatient admissions from the ED, which could aid to estimate hospitalization ward staffing needs have been devised [40]. However, these models do not provide sufficient advance warning [41] to take effective measures, aside from ambulance diversion and outbound patients transfers. Such models may be useful for *a posteriori* analyses, but cannot be applied for effective staff planning as this usually requires accurate medium-term (at least 2 weeks ahead) predictions. On the other hand, a large body of research has been dedicated to ED prediction with autoregressive time series models for medium and long-term forecasts [42]. Some of the previous studies did not make a distinction between train and test sets [43-46]. In other cases, model reliability was not tested upon sufficiently long or varied periods of time. Also, most often, monthly [47] or daily visits were predicted [43, 44, 48-51]. Daily predictions are of little use for realistic staff planning, since nursing personnel is usually organized in 8-hour shifts and patient arrival volume often varies dramatically between these shifts, as fewer patients seek ED care at night or during morning working hours than during the afternoon shift. Additionally, quantifications of overstaffing and understaffing

situations were lacking in previous research, although the stated purpose of most of the models was precisely staff level planning [42].

Previous models have usually relied on calendar variables such as day of the week, week number, month of the year, festivity (yes/no) [42]. Some authors found evidence of effects of weather-related factors for specific pathologies [52-54], however, these have been shown to have very little or no effect upon *total* ED visits and thus can be excluded from the analysis [42, 43]. Most importantly, weather-related variables most often cannot be obtained with sufficient accuracy, reliability and advance for ED personnel planning. Televised football games could have an effect on the usage of emergency care services. However, the few studies devoted to this topic in Spain showed that televised football and high-audience television programming in general did *not* seem to influence ED visits in a significant manner. In 2004, in Alicante (Comunidad Valenciana region, Spain), no statistically significant differences were observed between trauma ED visits during days with football matches and days without football matches. In this study, 46 days were compared, 14 of these were Saturdays with football matches [55]. In 2009 in Murcia (Murcia region, Spain) no statistically significant differences were found between days with high-audience televised events (audience of more than 6.000.000 spectators) and control days (without high-audience televised events). 48 days were compared in this study: 24 days with high-audience televised events and 24 control days. 19 of the 24 days with high-audience events were football matches [56]. We thus chose not to include football matches or televised events as independent variables.

2.2.2. Autoregressive integrated moving average and seasonal integrated autoregressive moving average models

Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models are extensions of the ARMA class models, which include more realistic dynamics, in particular, respectively, non stationarity in mean and seasonal behaviors [57].

Non-seasonal ARIMA models generally use the notation $ARIMA(p,d,q)$, where parameters p , d and q are non-negative integers. p is the order of the autoregressive model, d is the degree of differencing and q is the order of the moving-average model.

Time series often possess a seasonal component that repeats every S observations. In order to deal with seasonality, SARIMA models have been formulated [58]. The notation generally used for SARIMA models is $ARIMA(p,d,q) \times (P,D,Q)_S$. The seasonal part of the model consists of terms that are very similar to the non-seasonal components of the model, but involve backshifts of the seasonal period [59]. S refers to the number of periods in each season. P , D and Q refer to the autoregressive, differencing and moving-average terms for the seasonal part of the ARIMA model.

When a time series is characterized by nonstationarity in variance (heteroscedasticity), it may be removed with transformations such as Box-Cox [58, 60].

$$Z_t(\lambda) = \frac{X_t^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0$$

$$Z_t(\lambda) = \log(X_t) \text{ if } \lambda = 0$$

where X_t is the original time series, Z_t is the transformed time series and λ is a parameter .

λ can be determined with maximum likelihood estimates (most often, the first transformations to attempt are logarithmic transformations with $\lambda = 0$). A good value of λ should reduce heteroscedasticity and also make the process more similar to a Gaussian and also reduce the effects of possible outliers [60]. However, in practice, this transformation seldom fulfills the basic assumptions of linearity, normality and homoscedasticity *simultaneously* [61]. Also, complex transformations can significantly reduce the interpretability of results.

ARIMA-style models are severely constrained by their assumptions [62], and even in the late 1970s and early 1980s, it became clear that they were not adequate for many applications [63]. In large part due to the requirements of overcoming some of these limitations for financial time series, several parametric techniques have been proposed, such as autoregressive conditional heteroskedastic (ARCH) models [64], generalized ARCH models (GARCH) [65] or threshold autoregressive models. However, the adequacy of these models for *forecasting* purposes is contested [63].

We decided to include SARIMA models as a benchmark. ARIMA and SARIMA models can be generated with the open-source **forecast** package in R [66] for medium-sized datasets with automated parameter fitting. However, the **forecast** package is limited to time series with just *one level of seasonality*. To the best of our knowledge, alternative open source software packages capable of handling *several* levels of seasonality, such as X-13ARIMA-SEATS, are only capable of handling (very) short time series [67].

2.2.3. Machine learning time series models

A machine learning approach based on the machine learning methods for time series forecasting, is generally more flexible and powerful than the autoregressive time series modeling approach [68] since (i) stationarity is not required, (ii) any number of cyclical variables as well as non-linear terms can be easily included, (iii) autoregressive models do not allow the inclusion of long-lag terms while excluding shorter-lag terms, while the machine learning approach does, and (iv) any kind of algorithm with a continuous output can be used in the regression approach [62].

In time-series models lags refer to preceding values of the dependent variable. For instance, “lag 1” refers to the value of the dependent variable at $t-1$. “lag 2” refers to the value of the dependent variable at $t-2$. However, unlike ARIMA-style models, algorithms used in the generalized regression approach for time series models may include non-consecutive lags. For example, given “maximum lags = 5”, i.e. the introduction of a maximum of 5 lag terms in the model, a forecasting algorithm based on a generalized regression approach may ignore lags 1, 2, 3 and 5, but may use lag 4.

In the last two decades, machine learning-based time series models have drawn attention and have established themselves as serious contenders to classical statistical autoregressive models in the forecasting community [62, 69, 70]. ANNs [71], decision trees and nearest neighbor regression [72] have been successfully used for forecasting nonlinear time series. Also, SVR time series models have been previously used for many other problems such as financial forecasting [73], weather and environmental state prediction [74], as well as reliability forecasting [74]. SVR time series have been shown to be superior to ARMA forecasting techniques for similar problems [75].

2.2.3.1. Classification and regression trees, M5P and AMT

Classification and regression trees (CARTs) can be defined as predictive graphics or as models based on a tree of questions in the branches (or decision nodes) and sets of solutions in the leaves (or end nodes). CARTs can be used as white-box decision support systems and are usually simple to understand after a brief explanation, although interpretability decreases with the size of the tree [76].

The basic algorithm behind CARTs is to i) start with all variables in a group, ii) find the variable split that better separates the outcomes, iii) divide the groups into groups (*leaves*) onto that split (*node*), iv) within each split (*node*), find the variable split that better separates the outcomes, v) repeat steps iii) and iv) until the groups are too small or sufficiently homogeneous (*pure*). Some measures of *impurity* are: misclassification error, Gini impurity and deviance / information gain [77]. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered-discrete values, with prediction errors typically measured by the squared difference between the observed and predicted values [78].

The M5P is an example of a tree algorithm for regression problems and is based on the M5 algorithm [79]. Starting at the root node, at each node, a splitting decision is taken based on a test outcome. In the case of M5P, it is an error minimization test. The expected error is calculated as

$$\Delta error = stdev(S) - \sum_i \left(\frac{|S_i|}{|S|} stdev(S_i) \right)$$

where S is the set of instances passed to the node, $stdev(S)$ is its standard deviation, S_i is the subset of S resulting from splitting at the node with the i^{th} outcome for the test [80]. Splitting stops if values in a node exhibit low inner-class variation or the number of instances is less than a pre-defined minimum (M). M is a parameter that can be defined experimentally.

Then, pruning is applied and pruned inner nodes are transformed into regression planes. Finally, a smoothing procedure is used to avoid sharp discontinuities [80]. M5P tree time series have been used for forecasting directional changes in financial markets [81], solar irradiation [82] or traffic incidents [83].

Alternating Model Trees (AMT) is a recently developed ensemble algorithm with a single tree structure. AMT may be considered as an evolution of the M5P tree algorithm. In M5P *multivariate* regressions may be used on leaves, while, in AMT, only *univariate* (also called *simple*) regressions are allowed on leaves. However, while M5P employs the classical approach for growing and pruning decision trees, AMT relies on a boosting algorithm (additive regression with forward *stagewise* selection) for tree growth and cross-validation for tree size determination. The AMT algorithm has been shown to produce lower root relative squared error (RRSE) than M5P on some datasets [84] and has been used for the development of time series models for enterprise service-level performance indicators such as demand, throughput, average order-delivery time, quality of service, or end-to-end cost [85]. In the Weka implementation [86], AMT is determined by two parameters: I (number of iterations) and H (a shrinkage parameter). I is a positive integer and H is a decimal number bounded to the $(0, 1]$ interval. $H = 1$ means that no shrinkage is performed [84].

2.2.3.2. Support Vector Machine Regression

Support Vector Machines (SVMs) are based on sound theoretical principles and can be formulated as convex optimization problems [87]. Support vector machines (SVMs) are arguably the best known example of kernel learning algorithms [88]. For a two-class linearly separable learning task, the aim of SVMs is to find a hyperplane that can separate a set of samples into two classes with a maximal margin [87]. Intuitively, a margin can be defined as the amount of space, or separation, between the two classes as defined by a hyperplane. Geometrically, the margin corresponds to the shortest distance between the closest data points to any point on the hyperplane. However, in many real-world problems, it may be too rigid to require that all points are linearly separable, especially in many complex nonlinear classification problems [76]. In these cases the optimization problem is not solvable as-is [89]. There are two approaches to transform the problem into a solvable form. The first one is to relax the rigid inequalities in and thus lead to the so-called soft margin optimization [76]. Another method, which is more common, is to apply the “kernel trick” to linearize nonlinear problems [90].

Support Vector Regression (SVR) models are an adaptation of SVMs for regression problems [91, 92]. They can also be interpreted as a variation over traditional linear regression models. The latter use the least difference between predicted and observed values, while the former rely on the minimization of the generalized error bound, which usually results in better predictions. One of the most popular variations [76] of SVR is ε -SVR. ε -SVR models are defined by parameters C , ε , kernel type and kernel-specific parameters [91]. Parameter C is the “complexity”, and can be interpreted as a regularization parameter. Decreasing C may result in better results if there is evidence of overfitting, i.e. the model produces very good results on training data, but fails to generalize them, which produces poor results on new testing data. Conversely, if the model is

underfitting (not producing a good fit even on the training data) increasing C could improve results [93]. Parameter ε is the “tolerance termination criterion” and affects tolerance to noise in the data [92]. A kernel type defines the function used for solving the optimization problem defined in the SVR [94]. In the Weka implementation [95], a polynomial kernel is defined by parameter E , which is the exponent of the polynomial. $E = 1$ results in a linear kernel.

Support Vector Regressions have been used for a large variety of problems such as financial time series [73], benchmarking of promotional efficiency [96], the prediction of porosity in heterogeneous reservoirs [97], forecasting of wind turbine heat rates [98] or wind turbine power coefficient estimations [99], among many others.

2.2.4. Risks measurements in forecasting problems for ED nursing personnel allocation

The Mean Absolute Error (MAE) is a widely used and intuitive error measurement metric for time series, where absolute differences between observed (O) and predicted (P) values, obtained in N samples, are averaged:

$$MAE = \frac{\sum_i^N |O_i - P_i|}{N}$$

However, MAE by itself, and other commonly-used time series error metrics based on some form of absolute averages such as Mean Absolute Percentage Error (MAPE), Symmetric MAPE (sMAPE), or Root Mean Squared Error (RMSE) are insufficient for realistic risk estimations for ED personnel allocation in the ED, since understaffing and overstaffing do not have the same implications in ED management [100]. While overstaffing can be interpreted as a potential waste of resources, understaffing may produce clinical errors, adverse clinical outcomes [19], patients leaving without being seen, and staff burnout [12], among other consequences. Therefore, MAE must be complemented with other risk measurements for this particular problem, such as overstaffing and understaffing risks. Recommendations for clinical staffing levels state that one nurse should be available *for every four patients* in the ED *at all times* [20, 27, 30, 31]. These recommendations can be used to transform ED census levels into nursing staffing levels. Overstaffing and understaffing risks can be derived from these estimations. Staffing costs should also be taken into account. Given an estimated average cost per nurse and 8-hour shift, an estimation of a staffing allocation policy can be obtained by means of adding up the predicted nursing personnel costs for all 8-hour shifts [101].

2.3. Materials and Methods

2.3.1. Data source

The study setting was the Ramón y Cajal University Hospital of Madrid, a 1.100-bed public specialized care referral center. Anonymized ED episodes from years 2010 to 2014 were obtained from the HP-HIS central hospital information system. At the time of this study, this hospital did not possess a datawarehouse system. Thus, data extraction, validation and transformation had to be done with custom software. Functionality similar to that of a datawarehouse was achieved with a custom development based on QlikView. The resulting system behaved as a MOLAP (multidimensional online analytical processing), with query times at least an order of magnitude faster than similar systems based on traditional relational databases (such as MySQL, Oracle or SQL Server). Aggregated ED census levels were obtained following the schemes defined in **Figure 2**.

Since the information was aggregated, no permission from the hospital's Ethics Committee for Clinical Research was necessary.

2.3.2. Dependent variables: ED visits *versus* ED census levels

Previous research has often focused on ED *visit* forecasting. We also had developed a model with such an approach [101]. However, it is more accurate to use the number of patients at the ED at a given moment (i.e. ED census levels) for resource planning (**Figure 1**). The problem can be simplified considering that hour-by-hour forecasts are not needed. Instead, averages and maximums of census levels of each of the hours in a given 8-hour shift may be used (**Figure 2**). Nurse staffing levels should be adjusted to the *maximum* ED census needs in a given shift in order to comply with international and Spanish staffing recommendations, since nursing personnel is usually organized in 8-hour shifts and, as mentioned previously (**Section 2.2.4**), one nurse available for every four patients in the ED *at all times*.

Average ED census levels may be easier to forecast, may exhibit lower absolute errors, and also may result in lower personnel costs. Thus, ED managers may be tempted to use them for nursing personnel planning. We believe that this should *not* be done, and we found it relevant to generate *average* ED census forecasts and analyze the *risks* of this approach for nurse staffing planning. Also, even though *average* ED census forecasts should *not* be used for nurse staffing planning, they could be used for other purposes, such as clinical laboratory or radiological resource planning; as well as non-clinical personnel allocation (administrative personnel, security guards, etc).

The methodology we had developed for the forecasting of ED *visits* [101] could also be applied for the forecasting of ED census levels since, but with different dependent variables.

Figure 1. Emergency department visits and Emergency department census.

Figure 2. Maximum and average hourly census levels in 8-hour shifts.

Usually, data in ED information systems is stored in a format wherein each table row corresponds to one patient visit with a timestamp of entry to the ED and another timestamp of exit from the ED, among other variables. It's easy to see that obtaining census levels (i.e. the total number of patients at a given time) requires transforming the data. Perhaps for this reason, the forecasting of ED census levels has been less frequently addressed in previous literature than the forecasting of ED visits. QlikView offers an elegant solution to this data transformation problem, by means of *intervalmatch joins* [102], which allow combining a transaction table with another table of discrete events (such as a calendar table). Such operations are significantly slower and more laborious to program with classical relational databases (such as Oracle, MySQL, Postgres or Microsoft SQL Server) or with data management tools available in statistical software packages such as SAS, Stata, SPSS or R.

We calculated hour-by-hour census levels for the whole period of the study. Then, we calculated average *and* maximum of hourly ED census levels within each 8-hour shift (**Figure 2**).

2.3.3. Independent variables

Recall that in time-series models lags refer to preceding values of the dependent variable (**Section 2.2.3**). For instance, “lag 1” refers to the value of the dependent variable at $t-1$. “lag 2” refers to the value of the dependent variable at $t-2$. In the Weka implementation “Maximum lags = T” introduces “T lags” in the models. T could not be established *a priori* and we decided to consider several possibilities: 0, 1, 3, 21, 42, 84, 168 lags (each lag corresponding to one 8-hour shift).

In addition to lags, we considered the inclusion of additional information as exogenous independent variables:

- Month, day-of-week and shift (morning, afternoon and night shifts) were included in order to account for seasonality effects. These were always included together and are referred to as the “basic” *group* of exogenous variables.
- National, Regional and Madrid's municipal holidays were also initially included, since they could affect population movements and, hence, also ED visits and ED census levels. No difference was made between holiday types. We also decided to include “long weekends” or “bank holiday weekends” (colloquially called “*puentes*” in Spain). These could have a smaller effect on ED visits than that of conventional holidays and were included as different variables. Bank holiday weekends are defined in **Table 1**. Similarly to the “basic” *group* of exogenous variables, all the variables related to holidays (holiday, 1-day bank holiday weekend, 2-day bank holiday weekend) were always included together and are referred to as the “holidays” *group* of exogenous variables.

Day X: National / Regional / Municipal Holiday
Day X + 1: marked as “ 1-day *bank holiday weekend* ”
Day X + 2: Saturday

Day X: National / Regional / Municipal Holiday
Day X + 1: marked as “ 2-day *bank holiday weekend* ”
Day X + 2: marked as “ 2-day *bank holiday weekend* ”
Day X + 3: Saturday

Table 1. Definition of bank holiday weekend independent variables.

2.3.4. Performance measurements: underestimation risks, overestimation risks and costs

Recall that absolute error measurements were not sufficient for evaluating forecasting algorithms used for ED personnel allocation (**Section 2.2.4**). In our case, risks were defined according to **Table 2**, adjusted to *maximum* ED census levels within 8-hour shifts (**Figure 2**), considering that, as stated previously, one nurse should be available for every four patients *at all times* (according to national and international recommendations).

Severe understaffing was defined as an underestimation of more than 9 nurses between predicted staffing levels and required staffing levels. Medium understaffing was defined as an underestimation of 7 to 9 nurses. Low understaffing was defined as an underestimation of 4 to 6 nurses. Overstaffing was defined in a similar way (**Table 2**). These criteria were defined based on expert knowledge. We considered that, if the forecast error was within ± 3 nurses, the scenarios could be described as “very low understaffing” or “very low overstaffing”, and thus were not relevant. In the first scenario, nurses could assume an increased workload. In the second case, nurses not attending patients could temporarily perform other tasks. Therefore, these situations were not considered in further analyses.

Scenario	Description
4-6 nurse overstaffing	Minor overstaffing
7-9 nurse overstaffing	Medium overstaffing
> 9 nurse overstaffing	Major overstaffing
4-6 nurse understaffing	Minor understaffing
7-9 nurse understaffing	Medium understaffing
> 9 nurse understaffing	Major understaffing

Table 2. Overstaffing and understaffing scenarios for *maximum* ED census forecasts.

The emerging results were to be compared to the *usual resource allocation policies* in the Ramón y Cajal University Hospital (**Table 3**). Nursing personnel levels based on *maximum* ED census forecasts could be compared to “usual nursing personnel”, while other resources in the ED adjusted to *average* ED census levels could be compared to “equivalent patients” calculated as Usual nursing personnel x 4 patients / nurse.

Time shift	Usual nursing personnel (benchmark for <i>maximum</i> ED census levels)	Equivalent patients (benchmark for <i>average</i> ED census levels)
Mon – Fri Night Shift	35	140
Mon – Fri Morning Shift	40	160
Mon – Fri Afternoon Shift	40	160
Hol. / Sat / Sun Night Shift	35	140
Hol. / Sat / Sun Morning Shift	37	148
Hol. / Sat / Sun Afternoon Shift	37	148

Table 3. Usual resource allocation policies in the ED of the Ramón y Cajal University Hospital.

It is relevant to mention that the understaffing and overstaffing scenarios defined in **Table 2** can also be interpreted as percentage errors, considering the usual resource allocation policies defined in **Table 3**. For instance, given a staffing level of 40 nurses, a minor understaffing or overstaffing scenario is approximately equivalent to a percentage error range of 10% - 15%; a major understaffing or overstaffing scenario is equivalent to a percentage error higher than 22%; and a medium understaffing or overstaffing scenario is between the previous two. An interpretation based on percentage errors could serve to adapt the scenarios defined in **Table 2** to other ED settings.

To the best of our knowledge, no objective and publicly available criteria exist for defining a trade-off risk matrix for ED overstaffing, understaffing and costs in Spain (nor in other European countries). However, in the context of ED care, we believe that a human decision-maker would be most interested in avoiding understaffing *within budget restrictions*. We thus sought to explore different understaffing scenarios including calculations of personnel

expenses in the experiments. After consulting with the hospital’s Human Resources department, we assumed a gross cost of 125 euros per nurse and 8-hour shift, which included direct and indirect personnel costs. Finally, we defined our own optimization objectives based on expert recommendations: we aimed to reduce major and medium understaffing scenarios with a (monetary) cost increase of less than 10%. We also decided to include additional forecasting performance metrics that could be relevant for other decision-makers: MAE and estimations of overstaffing situations.

Recall that, although *average* ED census levels should not be used for nursing personnel planning, they could be used for other purposes, such as clinical laboratory or radiological resource planning; as well as non-clinical personnel allocation (administrative personnel, security guards, etc). Similarly to *maximum* ED census forecast error analyses, we define overestimation and underestimation risks for *average* ED census forecasts (**Table 4**).

Scenario	Description
16-27 patients overestimation	Minor overestimation
28-36 patients overestimation	Medium overestimation
> 36 patients overestimation	Major overestimation
16-27 patients underestimation	Minor underestimation
28-36 patients underestimation	Medium underestimation
> 36 patients underestimation	Major underestimation

Table 4. Overestimation and underestimation errors defined for *average* ED census forecasts.

For *average* ED census forecasts, “costs” were defined as the sum of predicted patients in all 8-hour shifts of the “macro test set” (**Section 2.3.5**). Depending on the specific problem, these could be transformed into monetary costs. As mentioned previously, we also aimed to reduce major and medium underestimation scenarios based on *average* ED census forecast with a cost increase of less than 10%.

2.3.5. Forecasting algorithm parameterization and custom time series evaluation scheme

We sought to evaluate different forecast horizons: $N = 1, 2, 4, 8$ weeks³, which is equivalent to 21 8-hour shifts, 42 8-hour shifts, 84 8-hour shifts and 168 8-hour shifts, respectively. From a managerial point of view, longer term forecasts are likely to be preferable; however, we decided to include several forecast horizons in order to evaluate their effect on forecasting performance. We also believe that forecasts should be generated with different forecast horizons since hospitals have varying flexibility for adjusting ED resources in advance. Even within the same hospital this flexibility may vary over time. Thus, expected underestimation risks, overestimation risks and costs should be known for several forecast horizons.

We considered three machine learning algorithms for the estimation of generalized regression time series: M5P, AMT and SVR with a polynomial kernel. We also decided to include a series of benchmarks: usual resource levels (**Table 3**), averages stratified by the three daily 8-hour shifts (stratified average), linear regression and SARIMA. Benchmark algorithms are described in more detail in **Section 2.3.5.2**. SARIMA parameters were automatically set and forecasts were produced with the **forecast** package available for the R statistical software [66]. Other benchmark algorithms required no parameterization.

For the machine learning algorithms (M5P, AMT and SVR), model parameterization was performed on validation sets, depending on the length of individual forecasts, in the manner described in **Table 5**. Parameterization could vary depending on the dependent variable (average or maximum ED census forecasts), the algorithm and the forecast horizon ($N = 1, 2, 4, 8$ weeks).

³ In this case “week” is any period of 7 consecutive days, not necessarily equal to *calendar* weeks.

Forecast horizon	Train dataset for parameterization	Validation dataset for parameterization
1 week (21 8-hour shifts)	January 1, 2010 – December 23, 2010	December 24, 2010 – December 31, 2010
2 weeks (42 8-hour shifts)	January 1, 2010 – December 16, 2010	December 17, 2010 – December 31, 2010
4 weeks (84 8-hour shifts)	January 1, 2010 – December 2, 2010	December 3, 2010 – December 31, 2010
8 weeks (168 8-hour shifts)	January 1, 2010 – November 4, 2010	November 5, 2010 – December 31, 2010

Table 5. Parameterization datasets

Once models were parameterized (for maximum and average ED census forecasts stratified by 8-hour shifts), we used a variation of a “sliding window evaluation approach”, similar to a previous methodological approach we had developed [101]. Model evaluation was performed simulating *what would have happened* if the predictions were made with the historical information available in the moment of each forecast, and subsequently calculating error metrics on the accumulation of predictions (**Figure 3**).

The initial train set was January 1, 2010 to December 31, 2010 and forecasts started on January 1, 2011. The train set was progressively increased. Test set lengths were fixed and defined by the forecast horizons of N weeks ($N = 1, 2, 4, 8$ weeks).

After the first forecast (that used the data from January 1, 2010 to December 31, 2010 as the train set), within each iteration,

- the train set was increased in N weeks;
- the models were recalculated on this larger train set (for M5P, AMT and SVR, algorithms were retrained using the parameter selections for each dependent variable, algorithm and forecast horizon that had been obtained on the validation sets defined in **Table 5**);
- a new forecast of N weeks ahead was performed.

The process stopped when the forecasts reached 31 December, 2014. The forecasts produced in all iterations were *concatenated* in “macro test sets”. In the end, “macro test sets” included forecasts from January 1, 2011 to December 31, 2014 (**Figure 3**).

Notice that “macro test sets” stored different forecasts depending on the dependent variable (average and maximum census forecasts), algorithm (models retrained on increasingly larger test sets) and forecast horizon ($N = 1, 2, 4, 8$ weeks).

For maximum ED census forecasts, Mean absolute errors (MAE), understaffing and overstaffing risks (as defined in **Table 2**) and forecasted personnel costs would be estimated for on the “macro test sets” (**Figure 3**). Similarly, for average ED census forecasts, MAE, underestimation and overestimation risks (**Table 4**) and resource costs were calculated on the “macro test sets”.

Let us use an example to further clarify this approach. Given $N = 2$ weeks, the first forecast was made for the period ranging from January 1, 2011 to January 14, 2011. A forecast was then made with models trained with the initial train set (January 1, 2010 to December 31, 2010). The results for this first prediction were stored in the corresponding “macro test set”. Then, a new prediction was made for the period ranging from January 15, 2011 to January 28, 2011 with models trained upon all data from January 1, 2010 to January 14, 2011. The results of this prediction were also stored in the corresponding “macro test set”. The process continued until forecasts reached December 31, 2014.

The usage of a growing train set was justified since, generally, the inclusion of more data (with an increasingly larger train set) improves a model if no overfitting occurs. Also, including recent data allows to adapt to changes in the hospital or in the characteristics of

the patient population ascribed to the hospital. If these conditions do not change, the inclusion of more data may not improve the results. If these conditions change strongly, it may be advisable to rely exclusively on more recent data.

Figure 3. Time series evaluation scheme.

Since results could change with a fixed-length test set, we also performed an evaluation with a variation of the scheme described so far, with a 1-year fixed-length train set, i.e. using the 365 days preceding the forecast as a train set in all cases (a more detailed description and results can be found in **Section 2.4.3**).

2.3.5.1. M5P, AMT and SVR forecasting algorithm parameterization

Algorithm parameterization was performed with the data partitions described in **Table 5**. Weka API implementations of the M5P, AMT and SVR algorithms were used in all cases.

Experiment planning was defined according to **Table 6**. As stated before, we decided to forecast average and maximum ED census levels with several forecast horizons: 1, 2, 4 and 8 weeks ahead, which is equivalent to 21 8-hour shifts, 42 8-hour shifts, 84 8-hour shifts and 168 8-hour shifts, respectively. We also decided to contemplate the possibilities of inclusion of holidays and bank holidays (**Table 1**) as independent exogenous variables. Maximum lags were also included in the experiments. As stated previously in **Section 2.3.3**, we decided to include 0, 1, 3, 21, 42, 84 and 168 maximum lags (each lag corresponding to one 8-hour shift).

Dependent variables = "average ED census", "maximum ED census"

Forecast horizons = 21, 42, 84, 168 8-hour shifts = 1 week, 2 weeks, 4 weeks, 8 weeks

Independent variables:

- Maximum lags = 0, 1, 3, 21, 42, 84, 168 8-hour shifts
- Combinations of exogenous variables = "basic", "basic + holidays"

where

"basic" exogenous variables: "month", "day of week" and "shift"

"holidays" exogenous variables: "holiday", "1-day *bank holiday weekend*", "2-day *bank holiday weekend*"

Table 6. ED census levels experiments: independent variables and forecast horizons.

Although our ultimate objective was to develop models with a realistic estimation of overstaffing and understaffing risks (**Table 2**), some events, such as *severe* understaffing or *severe* overstaffing situations were very unlikely to occur in the validation datasets defined in **Table 5**. Therefore, MAE was used as the metric for algorithm parameterization.

2.3.5.1.1. M5P parameterization

For the M5P tree algorithm (described in **Section 2.2.3.1**), besides the determination of the optimum number of maximum lags and combinations of exogenous variables, the value of parameter M (minimum number of instances in a node) was also set by following the parameter selection scheme described in **Table 7**. Parameter M was tested for values 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100.

```

for (Dependent variables to forecast) {
  for (Maximum lags) {
    for (Forecast horizons) {
      for (Combinations of exogenous variables) {
        for (Values of parameter M) {
          RunExperiment();
        }
      }
    }
  }
}

```

Table 7. M5P parametrization strategy.

The resulting optimal combinations of maximum lags, exogenous variable combinations and variations of parameter M for different fields and forecast horizons can be seen in **Table 8**.

Forecasted variable	Forecast horizon (8-hour shifts)	Min. MAE validation set	Parameter	Max. lags	Exogenous variables
Average census	21	8.6307	M=1	1	"basic" group of exogenous variables ("month", "day of week", "shift")
Average census	42	8.6609	M=45		
Average census	84	8.3691	M=5		
Average census	168	7.7105	M=5		
Maximum census	21	9.9438	M=20		
Maximum census	42	11.3639	M=20		
Maximum census	84	11.0891	M=20		
Maximum census	168	11.0936	M=25		

Table 8. Optimum M5P parametrization on the validation set.

Maximum census levels were better predicted using just the “basic” group of exogenous variables. In the case of *average census* forecasts, experimental results seemed to suggest the inclusion of holidays as exogenous variables. However, a closer examination revealed almost identical MAE metrics (less than 0.1% difference) on the validation sets for models with just the “basic” group of exogenous variables compared to models which used both “basic” and “holidays” groups of exogenous variables, for identical values of M and maximum lags. We thus decided to use only the “basic” group of exogenous variables in further experiments, following the parsimony principle.

Maximum lags = 1 produced the best results in all cases (**Table 8**). In order to assess the stability of this hypothesis, we decided to use the following methodology: (i) all models with the “basic” exogenous variables were selected, i.e. all possible models with the different combinations of parameter M and maximum lags for the “basic” group of exogenous variables; (ii) the selected models were then *averaged* by maximum lags for both prediction variables (*average census* and *maximum census*) and represented graphically. Our aim was to establish if the optimal value for maximum lags was indeed 1 for all combinations. Our hypothesis was confirmed (**Figure 4** and **Figure 5**).

For some forecast horizons, the difference between maximum lags = 1 and higher numbers was not large, however, a lower number of lags produced simpler models and thus reduced computational complexity. Therefore, maximum lags = 1 was preferable. With maximum lags = 1 the value of the dependent variable in the *preceding 8-hour shift* was included, while the remaining information of the time series could be captured by the basic group of exogenous variables ("month", "day of the week" and "shift").

Figure 4. Average of MAEs for "average ED census" forecasts with the "basic" group of exogenous variables and varying *maximum lags*.

Figure 5. Average of MAEs for "maximum ED census" with the "basic" group of exogenous variables and varying *maximum lags*.

Parameter M exhibited more variability. A linear effect on MAE as a function of M could not be assumed. Rather, a parabolic effect could be expected (high error for both low and high values of M), hence the usage of linear-assumption estimations with M as a parameter and MAE as the dependent variable, would have led to misleading results. However, once other factors could be set at fixed values (maximum lags = 1 and exogenous variables = “basic” for both maximum and average ED census forecasts), a graphical analysis in two dimensions was possible. Graphical representations of MAE on the train and validation sets with the four possible forecast horizons (21, 42, 84 and 168 8-hour shifts) can be found in **Figure 6** and in **Figure 7**. The X axis on these figures is parameter M of M5P tree, i.e. the *minimum* number of instances per leaf. In the case of average census forecasts, MAE was almost independent from parameter M. Considering the graphical analysis, it was set at M=5, as this value seemed to produce the lowest MAE for forecast horizons of 84 and 168 shifts and reasonably low MAE values for shorter forecast horizons. In the case of maximum ED census forecasts, considering the evolution of the error on both train and validation sets, the value M=20 seemed to produce a reasonably stable result in all scenarios (**Figure 7**).

The “valleys” in **Figure 6** and **Figure 7** are a logical consequence of the behavior of M5P. If the number of instances per leaf (parameter M) is too low, the model is likely to overfit (low MAE on the train set, high MAE on the validation set). If the number of instances per leaf is too high (parameter M), the tree becomes too general, with little or no difference between further increases in parameter M past a certain point.

Figure 6. Average of MAEs for "average ED census" with the "basic" group of exogenous variables, maximum lags = 1 and varying $parameter M$.

Figure 7. Average of MAEs for "maximum ED census" with the "basic" group of exogenous variables, maximum lags = 1 and varying $parameter M$.

2.3.5.1.2. AMT and SVR parameterization

Parameterization was repeated for SVR and AMT under the assumption of *independence* of exogenous variables and maximum lags from the parameters inherent to each algorithm.

We decided to follow a two-step parameterization strategy with fewer combinations and less computationally demanding experiments with Support Vector Regression (SVR) (introduced in **Section 2.2.3.2**) and Alternating Model Tree (AMT) (introduced in **Section 2.2.3.1**) time series (**Table 9**). The default parameter values were, respectively, I=10 (number of iterations = 10) and H=1.0 (no shrinkage) in the case of AMT and E=1 (linear kernel) and C=1.0 (complexity = 1.0) in the case of SVR.

Step 1:

```
for (Dependent variables to forecast) {
  for (Forecast horizons) {
    for (Maximum lags) {
      RunExperiment(DefaultParameters, Exogenous variables = "basic");
    }
    for (Combinations of exogenous variables) {
      RunExperiment(DefaultParameters, Maximum lags = 1);
    }
  } // end for (Forecast horizons)
}
```

Step 2:

```
for (Dependent variables to forecast) {
  for (Forecast horizons) {
    for (Values of parameter 01) {
      for (Values of parameter 02) {
        RunExperiment(Exogenous variables = "basic", Maximum lags = 1);
      }
    }
  } // end for (Forecast horizons)
}
```

Table 9. AMR and SVR parametrization strategy.

The results of “step 1” experiments (**Table 9**) confirmed the hypothesis of maximum lags=1 was an optimal value for *both* AMT and SVR, for average and maximum census forecasts.

Different combinations of exogenous variables produced negligible differences in MAE. Therefore, it was decided to use the “basic” group of exogenous variables, as in M5P.

For “step 2” experiments (**Table 9**) with the AMT algorithm, optimal values were established for the different forecast horizons and combinations of values $I = \{1, 2, 5, 10, 20, 40, 80, 160\}$ and $H = \{0.125, 0.25, 0.5, 0.75, 1.0\}$ (**Table 10**). However, these values could be unstable. We thus decided to represent all possibilities graphically in 3 dimensions (parameter I, parameter H and the resulting MAE) examining the performance on all validation sets and looking for "valleys", which could be assumed to offer certain guarantees of stability.

Forecasted variable	Forecast horizon (8-hour shifts)	Min. MAE validation set	Parameters	Max. lags	Exogenous variables
Average census	21	8.0282	I=5; H=1.0	1	"basic" group of exogenous variables ("month", "day of week", "shift")
Average census	42	8.2029	I=10;H=0.5		
Average census	84	8.4606	I=5; H=1.0		
Average census	168	7.9202	I=5; H=0.75		
Maximum census	21	9.4571	I=10;H=1.0		
Maximum census	42	10.5767	I=10;H=1.0		
Maximum census	84	11.5737	I=10;H=1.0		
Maximum census	168	11.3804	I=40;H=0.25		

Table 10. AMT initial parametrization.

Let us illustrate this with *average* census forecasts. Let us see an example with a forecast horizon of 84 shifts (**Figure 8**). Even though the optimal value was indeed $I=5$ and $H=1.0$ (marker in **Figure 8**), a graphical representation suggested that the area between $I=36$ and $I=60$ and $H=0.4$ and $H=0.7$ was slightly more *stable* (red zone in **Figure 8**). For this reason, we chose $H=0.5$ and $I=40$.

Figure 8. MAE on the validation set for the prediction of average ED census with a forecast horizon of 84 shifts (4 weeks ahead) for an AMT algorithm with variable parameters H and I ($H=0..1$; $I=0..160$). *The marker ($X=Parameter\ I\approx 5$; $Y=parameter\ H=1$; $Z=MAE\approx 8.4$) is set at an optimum which is deemed unstable. The red zone is assumed to be a stable optimum.*

In the case of maximum census forecast with a horizon of 168 shifts, optimal parameters were $I=40$ and $H=0.25$ (Table 10), as opposed to the optimal values obtained for all other cases ($I=10$ and $H=1.0$). After representing the function graphically, we decided that values $I=10$ and $H=1.0$ were also adequate. The final parameterization of AMT can be seen in Table 11.

Forecasted variable	Forecast horizon (shifts)	Min. MAE validation set	Parameters	Max. lags	Exogenous variables
Average census	21	8.0282	I=10; H=0.2	1	"basic" group of exogenous variables ("month", "day of week", "shift")
Average census	42	8.2029	I=40; H=0.1		
Average census	84	8.4606	I=40; H=0.5		
Average census	168	7.9202	I=40; H=0.9		
Maximum census	21	9.4571	I=10; H=1.0		
Maximum census	42	10.5767	I=10; H=1.0		
Maximum census	84	11.5737	I=10; H=1.0		
Maximum census	168	11.3804	I=10; H=1.0		

Table 11. AMT reviewed parametrization.
Differences with initial parametrization in bold.

For “step 2” experiments (**Table 9**) with the SVR algorithm, optimal values were established for the different forecast horizons with experiments with all combinations of values $E = \{1.0, 2.0, 3.0\}$ and $C = \{0.00625, 0.0125, 0.025, 0.05, 0.1, 1.0, 2.0, 4.0, 8.0, 16.0\}$.

Results were represented with 3-dimentional surfaces for all four forecast horizons (21, 42, 84 and 168 8-hour shifts). All had the same shape (MAE monotonically decreasing with the decrease of both parameters) for both average ED census levels and maximum ED census levels on the *validation* set, hence, only the 84-shift forecasts for maximum census levels are shown as an example in **Figure 9 B**. The effect on the train sets was the *opposite* (**Figure 9 A**, notice that the X and Y axes follow *opposite* directions to those of **Figure 9 B**): higher values of E resulted in a nearly perfect fit, and hence a lower value of MAE on the train set. However, given the results on the validation set (**Figure 9 B**) it was clear that models resulting from values of E higher than 1 failed to generalize.

The conclusion for all forecast horizons and both average and maximum census was the same: the best results on the validation sets were obtained with $E = 1.0$, i.e. a linear kernel. After setting E at this fixed value, the effect of a variation of parameter C could now be represented as two-dimensional graphics for all forecast horizons and reasonable values of parameter C for average census (**Figure 10**) and maximum census (**Figure 11**). Given these graphical representations, we decided to set $E = 1.0$ and $C = 0.1$, a parameterization that seemed to produce a consistent minimization of MAE, for all forecast horizons and both average and maximum census predictions, especially for longer term forecasts (168 8-hour shifts = 8 weeks).

(A)

(B)

Figure 9. MAE on the (A) train set and (B) validation set as a function of parameters C and E of an SVR time series for the prediction of *maximum census* and a forecast horizon of 84 shifts (4 weeks ahead).

Figure 10. MAE for a forecast of “average ED census” for an SVR algorithm with fixed parameter E=1.0 (linear kernel) and variable parameter C (regularization).

Figure 11. MAE for a forecast of maximum ED census for an SVR algorithm with fixed parameter $E=1.0$ (linear kernel) and variable parameter C (regularization).

2.3.5.2. Benchmarks: usual staffing levels, Stratified Average, linear regression and SARIMA

As stated previously, the algorithms defined in the previous section were compared to several benchmarks: (i) the usual resource allocation policies in the Ramón y Cajal University Hospital (**Table 3**); (ii) a “Stratified Average (SA)” model; (iii) ordinary least squares linear regression; (iv) Seasonal ARIMA (SARIMA) models.

The “Stratified Average” (SA) model was implemented with a calculation of mean values of average and maximum ED census levels for each one of the three daily 8-hour shifts. This calculation was performed using the train set available at each point. For example, for forecasts starting at January 1, 2010, average values stratified by the three daily 8-hour shifts were calculated for all the data starting from January 1, 2010 to December 31, 2010.

In the case of SARIMA, we decided to include models with 21 shifts (1-week seasonality) and 84 shifts (4-week seasonality). SARIMA parameters were automatically set and forecasts were produced with the **forecast** package available in R [66]. As mentioned previously (**Section 2.2.2**), **forecast** allowed only one level of seasonality in SARIMA models.

2.3.5.3. Evaluation of results

Experiments were executed following the scheme we had previously described (**Section 2.3.5**, and, specifically, **Figure 3**). MAEs for *maximum* ED census levels, overstaffing and understaffing risks (**Table 2**), as well as estimated monetary costs, were calculated for all algorithms and all forecast horizons. MAEs for *average* ED census levels, overestimation and underestimation risks (**Table 4**), as well as estimated monetary costs, were also calculated for all algorithms and all forecast horizons.

Since the sample sizes were sufficiently large in all cases, 95% confidence intervals (CIs) for overstaffing and understaffing situations were estimated with the Wald confidence interval for a proportion. 95% CIs for overestimation and underestimation scenarios were calculated with the same procedure.

In the case of *maximum* ED census forecasts, monetary costs of nursing personnel (assuming a gross average cost of 125 euros per 8-hour shift per nurse) were obtained on the “macro test sets” resulting from the experiments performed with each algorithm and forecast horizon. Given that the “macro test set” for each algorithm and forecast horizon stored four years of forecasts (**Section 2.3.5**), the total predicted personnel costs for each algorithm were divided by four in order to obtain approximate yearly costs. In the case of *average* census forecasts, costs were defined as the sum of patients predicted by each algorithm on each 8-hour shift divided by four (thus also obtaining yearly costs, although in different units).

MAE was also obtained on the accumulation of forecasts on the “macro test set” corresponding to each dependent variable, algorithm and forecast horizon.

2.4. Results

2.4.1. Maximum ED census levels

Detailed results for staffing levels resulting from *maximum* census forecasts can be found in **Table 12**. For clarity purposes, point estimates (i.e. estimates without confidence intervals) of algorithm performance in terms of understaffing risks and overstaffing risks (defined in **Table 2**) are shown in **Figure 12** and **Figure 13**, respectively. Average yearly costs of nurse allocation policies based on different algorithms are shown in **Figure 14**.

Remarkably, major, medium and minor understaffing risks were almost unaffected by the forecast horizon in the cases of linear regression, SVR, AMT and M5P (**Figure 12**). As expected, the risks of understaffing of SA and usual staffing levels (**Table 3**) were also unaffected by the forecast horizon. In the case of major understaffing (>9 nurses) and medium understaffing (7-9 nurses), the performance of SARIMA models declined with the increase of length of the forecast horizons (**Figure 12**).

Recall that our objective was to reduce major and medium understaffing situations with a cost increase of less than 10%, compared to usual resource allocation policies.

M5P was the best choice for the reduction of major and medium understaffing scenarios for all forecast horizons. Compared to the *usual nurse staffing levels* (**Table 3**), nursing personnel planning with M5P (**Figure 12** and **Table 12**) could reduce major understaffing (>9 nurses) risks more than 10-fold (a reduction to ~1% with M5P compared to ~13% obtained with the usual nurse staffing levels); and could reduce medium understaffing (7-9 nurses) risks approximately 3-fold (a reduction to ~3% with M5P compared to ~10% obtained with the usual staffing levels). However, the usage of M5P would have also implied

personnel costs of approximately 4,091,000 € - 4,098,000 € per year (depending on the forecast horizon), higher than the usual staffing costs, which were estimated to be 3,863,812.50 € per year (**Table 12** and **Figure 14**), i.e. ~5% - 6.1% increases in personnel costs (depending on the forecast horizon). This fulfilled our objective of significant reduction of understaffing with a less than 10% cost increase.

Although M5P was the best choice for understaffing avoidance, its 95% confidence intervals (CIs) overlapped with those of other algorithms in some scenarios (**Table 12** and **Figure 12**). For major understaffing (>9 nurses), M5P's 95% CIs overlapped with those of AMT, SVR and linear regression (**Table 12**) for 1-week, 4-week and 8-week forecast horizons. For 2-week forecast horizons, M5P's 95% CIs overlapped with AMT and linear regression. However, for medium understaffing (7-9 nurses) avoidance M5P was the best algorithm, with statistically significant differences with the rest of the algorithms excepting AMT for 1-week, 2-week and 8-week forecast horizons. In the case of 4-week forecast horizons and medium understaffing, M5P exhibited statistically significant differences with *all* other algorithms, including AMT. In the case of minor understaffing (4-6 nurses), M5P's 95% CIs overlapped with those of SA, AMT and SARIMA (S=1 week) for all forecast horizons, and also with those of SARIMA (S=4 weeks) for forecast horizons of 2 and 8 weeks.

In the case of maximum ED census forecasts there was no clear winner for the avoidance of *overstaffing* situations (**Table 12** and **Figure 13**). For *major* (>9 nurses) and *medium* (7-9 nurses) overstaffing there were no statistically significant differences between linear regression, SVR, AMT and M5P. However, these three exhibited lower risks of major and medium overstaffing than SA, usual staffing, SARIMA (S=1 week) and SARIMA (S=4 weeks), for all forecast horizons. For *minor overstaffing* (4-6 nurses), the differences between different algorithms were small, for all forecast horizons.

Algorithm	Forecast horizon	MAE	% Major (>9 nurses) understaffing 95% CI	% Medium (7-9 nurses) understaffing 95% CI	% Minor (4-6 nurses) understaffing 95% CI	% Minor (4-6 nurses) overstaffing 95% CI	% Medium (7-9 nurses) overstaffing 95% CI	% Major (>9 nurses) overstaffing 95% CI	Estimated yearly monetary cost (€)
Usual staffing	1 week	22.73	(12.38%, 14.39%)	(9.38%, 11.18%)	(13.31%, 15.38%)	(11.93%, 13.92%)	(6.52%, 8.06%)	(3.65%, 4.85%)	3,863,812.50
SA	1 week	18.69	(5.00%, 6.37%)	(6.34%, 7.86%)	(11.85%, 13.83%)	(14.44%, 16.58%)	(8.52%, 10.25%)	(3.84%, 5.07%)	4,117,718.75
SARIMA (S=1 week)	1 week	17.04	(3.08%, 4.19%)	(5.63%, 7.07%)	(11.34%, 13.28%)	(15.73%, 17.94%)	(7.71%, 9.37%)	(3.01%, 4.11%)	4,140,343.75
SARIMA (S=4 weeks)	1 week	17.98	(3.99%, 5.24%)	(6.93%, 8.51%)	(12.40%, 14.42%)	(13.93%, 16.04%)	(7.76%, 9.42%)	(3.18%, 4.31%)	4,085,781.25
Linear reg.	1 week	12.35	(1.03%, 1.71%)	(4.29%, 5.58%)	(14.28%, 16.42%)	(10.06%, 11.91%)	(2.57%, 3.60%)	(0.35%, 0.79%)	4,013,718.75
SVR	1 week	12.30	(1.11%, 1.82%)	(4.36%, 5.65%)	(14.55%, 16.70%)	(9.91%, 11.75%)	(2.53%, 3.55%)	(0.22%, 0.60%)	4,004,062.50
AMT	1 week	11.45	(0.65%, 1.22%)	(2.99%, 4.09%)	(11.47%, 13.43%)	(10.21%, 12.08%)	(2.32%, 3.30%)	(0.33%, 0.77%)	4,060,937.50
M5P	1 week	11.38	(0.57%, 1.12%)	(2.38%, 3.37%)	(10.46%, 12.34%)	(11.16%, 13.10%)	(2.61%, 3.65%)	(0.48%, 0.98%)	4,096,093.75
Usual staffing	2 weeks	22.73	(12.38%, 14.39%)	(9.38%, 11.18%)	(13.31%, 15.38%)	(11.93%, 13.92%)	(6.52%, 8.06%)	(3.65%, 4.85%)	3,863,812.50
SA	2 weeks	18.72	(5.02%, 6.40%)	(6.32%, 7.84%)	(11.85%, 13.83%)	(14.46%, 16.61%)	(8.52%, 10.25%)	(3.84%, 5.07%)	4,117,500.00
SARIMA (S=1 week)	2 weeks	18.17	(3.69%, 4.89%)	(6.32%, 7.84%)	(11.03%, 12.95%)	(15.33%, 17.52%)	(7.61%, 9.25%)	(4.68%, 6.01%)	4,153,687.50
SARIMA (S=4 weeks)	2 weeks	20.09	(5.56%, 7.00%)	(8.04%, 9.73%)	(12.33%, 14.35%)	(12.62%, 14.65%)	(8.13%, 9.82%)	(5.80%, 7.26%)	4,089,468.75
Linear reg.	2 weeks	12.50	(1.09%, 1.79%)	(4.40%, 5.70%)	(14.28%, 16.42%)	(9.97%, 11.82%)	(2.72%, 3.77%)	(0.40%, 0.88%)	4,012,812.50
SVR	2 weeks	12.37	(1.15%, 1.87%)	(4.44%, 5.74%)	(14.48%, 16.63%)	(9.82%, 11.65%)	(2.59%, 3.62%)	(0.24%, 0.63%)	4,001,562.50
AMT	2 weeks	11.58	(0.65%, 1.22%)	(3.31%, 4.46%)	(11.87%, 13.85%)	(10.46%, 12.34%)	(2.22%, 3.18%)	(0.37%, 0.82%)	4,055,093.75
M5P	2 weeks	11.55	(0.59%, 1.14%)	(2.53%, 3.55%)	(10.54%, 12.43%)	(11.40%, 13.36%)	(2.63%, 3.67%)	(0.59%, 1.14%)	4,098,062.50
Usual staffing	4 weeks	22.73	(12.38%, 14.39%)	(9.38%, 11.18%)	(13.31%, 15.38%)	(11.93%, 13.92%)	(6.52%, 8.06%)	(3.65%, 4.85%)	3,863,812.50
SA	4 weeks	18.78	(5.11%, 6.49%)	(6.32%, 7.84%)	(11.80%, 13.78%)	(14.48%, 16.63%)	(8.52%, 10.25%)	(3.82%, 5.04%)	4,115,437.50
SARIMA (S=1 week)	4 weeks	20.10	(6.19%, 7.70%)	(7.56%, 9.20%)	(11.74%, 13.71%)	(13.57%, 15.67%)	(7.50%, 9.13%)	(5.67%, 7.12%)	4,089,906.25
SARIMA (S=4 weeks)	4 weeks	22.61	(8.48%, 10.20%)	(9.20%, 10.99%)	(12.71%, 14.75%)	(11.29%, 13.24%)	(8.72%, 10.47%)	(7.50%, 9.13%)	4,048,031.25
Linear reg.	4 weeks	12.61	(1.15%, 1.87%)	(4.83%, 6.18%)	(14.41%, 16.56%)	(10.04%, 11.89%)	(2.53%, 3.55%)	(0.37%, 0.82%)	4,001,843.75
SVR	4 weeks	12.47	(1.23%, 1.97%)	(4.53%, 5.84%)	(15.10%, 17.29%)	(9.73%, 11.56%)	(2.53%, 3.55%)	(0.24%, 0.63%)	3,992,406.25
AMT	4 weeks	11.80	(0.91%, 1.56%)	(3.78%, 4.99%)	(11.89%, 13.87%)	(10.35%, 12.22%)	(2.28%, 3.25%)	(0.35%, 0.79%)	4,043,812.50
M5P	4 weeks	11.71	(0.73%, 1.33%)	(2.70%, 3.74%)	(10.74%, 12.65%)	(11.36%, 13.31%)	(2.76%, 3.82%)	(0.61%, 1.17%)	4,095,781.25
Usual staffing	8 weeks	22.73	(12.38%, 14.39%)	(9.38%, 11.18%)	(13.31%, 15.38%)	(11.93%, 13.92%)	(6.52%, 8.06%)	(3.65%, 4.85%)	3,863,812.50
SA	8 weeks	18.90	(5.11%, 6.49%)	(6.30%, 7.82%)	(11.69%, 13.66%)	(14.46%, 16.61%)	(8.57%, 10.30%)	(3.95%, 5.19%)	4,118,062.50
SARIMA (S=1 week)	8 weeks	22.31	(9.86%, 11.70%)	(9.01%, 10.77%)	(12.16%, 14.16%)	(11.93%, 13.92%)	(7.04%, 8.63%)	(6.02%, 7.50%)	3,982,593.75
SARIMA (S=4 weeks)	8 weeks	24.46	(10.96%, 12.88%)	(9.58%, 11.39%)	(11.69%, 13.66%)	(11.25%, 13.19%)	(8.76%, 10.51%)	(8.76%, 10.51%)	4,028,062.50
Linear reg.	8 weeks	12.86	(1.33%, 2.10%)	(5.37%, 6.78%)	(15.33%, 17.52%)	(9.40%, 11.20%)	(2.55%, 3.57%)	(0.35%, 0.79%)	3,981,125.00
SVR	8 weeks	12.63	(1.29%, 2.05%)	(5.04%, 6.42%)	(15.37%, 17.57%)	(9.49%, 11.30%)	(2.53%, 3.55%)	(0.24%, 0.63%)	3,981,906.25
AMT	8 weeks	11.79	(1.07%, 1.77%)	(3.35%, 4.50%)	(11.40%, 13.36%)	(10.10%, 11.96%)	(2.53%, 3.55%)	(0.38%, 0.85%)	4,051,031.25
M5P	8 weeks	11.90	(0.77%, 1.38%)	(2.93%, 4.01%)	(11.16%, 13.10%)	(11.56%, 13.52%)	(2.93%, 4.01%)	(0.57%, 1.12%)	4,091,937.50

Table 12. Results on the macro test set (4 years) with algorithms for maximum ED census. 95% CI overlapping with those of M5P are highlighted in bold.

Figure 12. Estimated risks of understaffing (calculated upon *maximum* ED census forecasts) for different algorithms and forecast horizons.

Figure 13. Estimated risks of overstaffing (calculated upon *maximum* ED census forecasts) for different algorithms and forecast horizons.

Figure 14. Estimated yearly nursing personnel costs (calculated upon *maximum* ED census forecasts) for different algorithms and forecast horizons.

2.4.2. Average ED census levels

As stated previously (**Section 2.3.2**), we believe that *average* ED census forecasts should *not* be used for nursing personnel allocation. An analysis of errors, risks and costs if *average* ED census forecasts were used for nursing personnel allocation when the actually needed number had to be established according to *maximum* ED census can be found in **Appendix A**. As a summary, in this scenario understaffing situations are reduced most with the *usual staffing levels*. However, *usual staffing levels* also imply a much higher incidence of overstaffing situations and higher estimated yearly costs than other algorithms.

Recall that average ED census levels could be used for other purposes (**Section 2.3.2**), such as clinical laboratory or radiological resource planning; as well as non-clinical personnel allocation (administrative personnel, security guards, etc). We thus had decided to include an analysis of average ED census forecasts with the overestimation and underestimation situations defined in **Table 4**.

In this case, once again, M5P was the best choice for the reduction of major and medium underestimation risks, for all forecast horizons (**Table 13** and **Figure 15**). In this case, the *usual resource allocation policies* (“equivalent patients” on **Table 3**) implied low risks of major and medium underestimation (~2% risk of major underestimation and ~3.7% risk of medium underestimation). Although M5P produced statistically significant reductions in major and medium underestimation compared to *the usual resource allocation policies* (more than 10-fold reduction in risks of major underestimation and approximately 3-fold reduction in risks of medium underestimation), these results were less relevant from a practical point of view than in the case of *maximum* ED census forecasts (described in the previous section).

On the other hand, most importantly, in this case (*average* ED census forecasts), M5P led to a cost *reduction* of more than 15% compared to usual resource allocation policies (**Table 13**). The objective of reducing major and medium underestimation risks was fulfilled since the differences between M5P and usual resource allocation policies were statistically significant, as CIs did not overlap, for all forecast horizons; even though the absolute differences were not large.

M5P also implied very large decreases in major, medium and minor *overestimation* risks compared to usual resource allocation policies (**Table 13** and **Figure 16**). With M5P, major overestimation risks were reduced more than 150-fold and medium overestimation risks were reduced more than 15-fold, compared to usual resource allocation policies. No relevant reduction was achieved in the case of minor overstaffing risks. In these cases, similar results in terms of reduction of risks of major and medium overestimation could be achieved with linear regression, SVR and AMT for all forecast horizons.

2.4.3. Evaluation with fixed-length train window

We repeated all experiments, with SVR, AMT and M5P having the same parameterization (**Section 2.3.5.1**) but with train sets of fixed 1-year length. That is, instead of using increasing train sets on the scheme described in **Figure 3**, the train sets had a fixed length composed of 1 year of data preceding the forecast. For example, for the forecasts starting at January 15, 2011, the preceding 365 days were used as a train set (i.e. data from January 14, 2010 to January 14, 2011, but *not* data from January 1, 2010 to January 13, 2010).

Compared to the increasing train sets approach, for both average and maximum ED census forecasts, results were generally *slightly worse* for major and medium underestimation risks for SARIMA (S=1 week), SARIMA (S=4 weeks), linear regression, SVR, AMT and M5P (example: for M5P major and medium understaffing risks increased 0.5%-2% and costs varied 1.5%-2.5%). Results for *usual resource allocation policies* were identical (since these were unaffected by the size of the train sets), and results for SA were very similar. No statistically significant *improvements* in terms of major and minor underestimation risks were obtained.

Algorithm	Forecast horizon	MAE	% Major (>36 patients) underestimation 95% CI	% Medium (28-36 patients) underestimation 95% CI	% Minor (16-27 patients) underestimation 95% CI	% Minor (16-27 patients) overestimation 95% CI	% Medium (28-36 patients) overestimation 95% CI	% Major (>36 patients) overestimation 95% CI	Estimated yearly cost (sum of predicted patients)
Usual resource allocation	1 week	26,17	(1.84%, 2.73%)	(3.16%, 4.28%)	(6.02%, 7.50%)	(15.61%, 17.83%)	(13.59%, 15.69%)	(23.10%, 25.64%)	123,774.00
SA	1 week	15,75	(2.97%, 4.06%)	(4.74%, 6.08%)	(9.07%, 10.85%)	(17.84%, 20.17%)	(6.76%, 8.32%)	(1.78%, 2.65%)	107,244.57
SARIMA (S=1 week)	1 week	13,44	(1.59%, 2.43%)	(4.42%, 5.72%)	(10.30%, 12.17%)	(15.04%, 17.22%)	(3.71%, 4.92%)	(0.79%, 1.40%)	105,870.85
SARIMA (S=4 weeks)	1 week	16,04	(3.06%, 4.16%)	(5.61%, 7.05%)	(12.91%, 14.96%)	(15.15%, 17.33%)	(6.78%, 8.34%)	(1.70%, 2.55%)	105,258.84
Linear reg.	1 week	9,35	(0.01%, 0.21%)	(1.49%, 2.30%)	(10.28%, 12.15%)	(8.24%, 9.94%)	(0.59%, 1.14%)	(0.01%, 0.21%)	104,163.34
SVR	1 week	9,39	(0.00%, 0.18%)	(1.68%, 2.53%)	(11.60%, 13.57%)	(6.95%, 8.53%)	(0.46%, 0.96%)	(0.04%, 0.28%)	103,080.88
AMT	1 week	11,52	(0.59%, 1.14%)	(3.50%, 4.68%)	(11.80%, 13.78%)	(13.42%, 15.50%)	(1.09%, 1.79%)	(0.01%, 0.21%)	104,550.72
M5P	1 week	8,36	(0.00%, 0.18%)	(0.63%, 1.20%)	(7.00%, 8.58%)	(7.89%, 9.56%)	(0.55%, 1.09%)	(0.01%, 0.21%)	105,074.66
Usual resource allocation	2 weeks	26,17	(1.84%, 2.73%)	(3.16%, 4.28%)	(6.02%, 7.50%)	(15.61%, 17.83%)	(13.59%, 15.69%)	(23.10%, 25.64%)	123,774.00
SA	2 weeks	15,79	(3.01%, 4.11%)	(4.74%, 6.08%)	(9.01%, 10.77%)	(17.84%, 20.17%)	(6.84%, 8.42%)	(1.82%, 2.70%)	107,245.08
SARIMA (S=1 week)	2 weeks	14,95	(2.15%, 3.10%)	(5.56%, 7.00%)	(11.05%, 12.98%)	(14.81%, 16.98%)	(5.30%, 6.71%)	(1.39%, 2.17%)	105,722.73
SARIMA (S=4 weeks)	2 weeks	18,78	(4.76%, 6.11%)	(6.67%, 8.22%)	(13.22%, 15.29%)	(13.26%, 15.34%)	(9.09%, 10.87%)	(4.08%, 5.33%)	105,289.59
Linear reg.	2 weeks	9,48	(0.01%, 0.21%)	(1.49%, 2.30%)	(10.90%, 12.81%)	(8.48%, 10.20%)	(0.67%, 1.25%)	(0.04%, 0.28%)	104,118.78
SVR	2 weeks	9,47	(0.00%, 0.18%)	(1.72%, 2.58%)	(12.04%, 14.04%)	(7.06%, 8.65%)	(0.48%, 0.98%)	(0.06%, 0.31%)	103,008.27
AMT	2 weeks	9,34	(0.07%, 0.34%)	(1.84%, 2.73%)	(9.29%, 11.08%)	(8.17%, 9.87%)	(0.40%, 0.88%)	(0.03%, 0.25%)	104,585.72
M5P	2 weeks	8,45	(0.00%, 0.18%)	(0.77%, 1.38%)	(7.08%, 8.68%)	(7.98%, 9.66%)	(0.55%, 1.09%)	(0.04%, 0.28%)	105,120.61
Usual resource allocation	4 weeks	26,17	(1.84%, 2.73%)	(3.16%, 4.28%)	(6.02%, 7.50%)	(15.61%, 17.83%)	(13.59%, 15.69%)	(23.10%, 25.64%)	123,774.00
SA	4 weeks	15,85	(3.16%, 4.28%)	(4.79%, 6.13%)	(9.01%, 10.77%)	(17.69%, 20.00%)	(6.97%, 8.56%)	(1.82%, 2.70%)	107,233.56
SARIMA (S=1 week)	4 weeks	17,04	(3.71%, 4.92%)	(7.26%, 8.87%)	(11.54%, 13.50%)	(14.99%, 17.17%)	(6.52%, 8.06%)	(2.42%, 3.42%)	104,731.24
SARIMA (S=4 weeks)	4 weeks	21,98	(7.87%, 9.54%)	(8.44%, 10.16%)	(13.22%, 15.29%)	(11.27%, 13.21%)	(9.51%, 11.32%)	(7.23%, 8.85%)	104,062.77
Linear reg.	4 weeks	9,55	(0.01%, 0.21%)	(1.55%, 2.38%)	(10.99%, 12.91%)	(8.39%, 10.11%)	(0.67%, 1.25%)	(0.03%, 0.25%)	103,924.69
SVR	4 weeks	9,54	(0.00%, 0.18%)	(1.88%, 2.78%)	(12.16%, 14.16%)	(7.06%, 8.65%)	(0.46%, 0.96%)	(0.04%, 0.28%)	102,842.48
AMT	4 weeks	8,71	(0.46%, 0.96%)	(0.85%, 1.48%)	(8.44%, 10.16%)	(6.76%, 8.32%)	(0.33%, 0.77%)	(0.04%, 0.28%)	104,018.20
M5P	4 weeks	8,52	(0.01%, 0.21%)	(0.73%, 1.33%)	(7.37%, 8.99%)	(7.91%, 9.59%)	(0.65%, 1.22%)	(0.03%, 0.25%)	105,057.29
Usual resource allocation	8 weeks	26,17	(1.84%, 2.73%)	(3.16%, 4.28%)	(6.02%, 7.50%)	(15.61%, 17.83%)	(13.59%, 15.69%)	(23.10%, 25.64%)	123,774.00
SA	8 weeks	15,96	(3.29%, 4.43%)	(4.83%, 6.18%)	(9.12%, 10.89%)	(17.62%, 19.93%)	(6.97%, 8.56%)	(1.82%, 2.70%)	107,229.84
SARIMA (S=1 week)	8 weeks	20,42	(8.28%, 9.99%)	(8.52%, 10.25%)	(11.85%, 13.83%)	(12.82%, 14.86%)	(7.02%, 8.61%)	(4.44%, 5.74%)	101,827.22
SARIMA (S=4 weeks)	8 weeks	24,00	(9.97%, 11.82%)	(9.29%, 11.08%)	(12.69%, 14.72%)	(10.59%, 12.48%)	(9.97%, 11.82%)	(8.74%, 10.49%)	103,481.42
Linear reg.	8 weeks	9,72	(0.04%, 0.28%)	(1.76%, 2.63%)	(11.91%, 13.90%)	(8.76%, 10.51%)	(0.50%, 1.01%)	(0.01%, 0.21%)	103,519.48
SVR	8 weeks	9,69	(0.00%, 0.18%)	(2.03%, 2.95%)	(13.00%, 15.05%)	(7.06%, 8.65%)	(0.38%, 0.85%)	(0.04%, 0.28%)	102,521.61
AMT	8 weeks	9,50	(0.87%, 1.51%)	(1.23%, 1.97%)	(9.47%, 11.27%)	(6.80%, 8.37%)	(0.54%, 1.06%)	(0.14%, 0.46%)	103,166.18
M5P	8 weeks	8,71	(0.10%, 0.40%)	(0.97%, 1.64%)	(7.69%, 9.35%)	(8.00%, 9.68%)	(0.61%, 1.17%)	(0.03%, 0.25%)	104,830.78

Table 13. Results on the macro test set (4 years) with algorithms for average ED census. 95% CIs overlapping with those of M5P highlighted in bold.

Figure 15. Estimated risks of underestimation (calculated upon *average* ED census forecasts) for different algorithms and forecast horizons.

Figure 16. Estimated risks of overestimation (calculated upon *average* ED census forecasts) for different algorithms and forecast horizons.

2.5. Discussion

Overall, M5P was the best choice for both *maximum* ED census forecasts and *average* ED census forecasts. It outperformed or at least matched the performance of the benchmark algorithms defined in **Section 2.3.5.2**: usual staffing levels, SA, linear regression and SARIMA models. The performance of M5P held for very long forecast horizons (including 8 week ahead forecasts), allowing for resource allocation in advance.

In the case of *maximum* ED census forecasts, M5P fulfilled the optimization objective of significantly reducing major and medium understaffing risks with a cost increase of less than 10%. Compared to the usual staffing levels, personnel planning with M5P could reduce major understaffing (>9 nurse) risks more than 10-fold (a reduction to ~1% with M5P compared to ~13% obtained with the usual nurse staffing levels); and could reduce medium understaffing (7-9 nurses) risks approximately 3-fold (a reduction to ~3% with M5P compared to ~10% obtained with the usual staffing levels). The usage of M5P also implied 5% - 6.1% increases in estimated nursing personnel costs (compared to the usual staffing levels), which are acceptable given the large understaffing risk reductions. A 5% - 6.1% yearly cost increase is equivalent to ~196,000€ - 236,000€. To put this in perspective, in Spain, the Ramón y Cajal University Hospital has a *yearly* budget well over 400,000,000€. We believe that such investments are clearly worthwhile since they could potentially improve user satisfaction, reduce clinical errors (which could lead to onerous legal, financial and reputational costs), streamline ED operations and reduce nursing staff burnout.

In the case of *average* ED census forecasts, M5P also fulfilled the optimization objective of significantly reducing major and medium underestimation risks with a cost increase of less than 10%. In fact a statistically significant reduction was achieved with a more than 15% cost *decrease*. On the other hand, although relative reductions of risks were large (more

than 10-fold reduction in major underestimation and approximately 3-fold reduction in medium underestimation; both compared to the usual resource allocation policies), the usual resource allocation policies already had low risks of major and medium underestimation ($\sim 2\%$ risk of major underestimation and $\sim 3.7\%$ risk of medium underestimation).

For both forecasting problems (*average* ED census forecasting and *maximum* ED census forecasting) parameterization experiments suggested that the inclusion of holidays as exogenous variables did not improve the results, and that lags were not particularly relevant (parameterization experiments suggested the inclusion of a maximum of 1 lag). Autoregressive SARIMA models performed poorly, most likely due to the difficulty of establishing a coherent autoregressive structure. Remarkably, the length of forecast horizons did not seem to affect performance in a particularly negative manner. The relatively simple M5P regression tree algorithm produced the best results. All this suggests that ad hoc algorithms based on sophisticated averaging (which should probably include month, day of week and shift, as well as *interactions* between them) could achieve similar results. However, the main advantage of the time-series models based on machine learning algorithms that we have developed is that these models can be generated and regenerated automatically (once reasonable parameters are chosen for each model).

Experiments with a fixed-length train window of 1 year (**Section 2.4.3**) produced slightly worse results in terms of risks of major and medium underestimation, but variations were not large. This suggests that the problem did not vary strongly in these years and that acceptable results could be achieved with just 1 year of historical information. We believe that periodic re-parameterization experiments (for example, at each new forecast) were unlikely to improve results in this case.

Even though MAE is a simple metric, in this case, the best algorithm could also have been choosing the one with the lowest MAE. However, understaffing, overstaffing, underestimation and overestimation risks illustrated the *asymmetry* of results (overestimation risks and underestimation risks) in some cases and provided relevant information for algorithm choice, particularly, for analyzing the shortcomings of some benchmark algorithms.

Our results highlight the difficulty healthcare providers may face if nursing personnel standards must be enforced *at all times*. Since all understaffing situations are almost impossible to avoid, perhaps, *acceptable* understaffing avoidance scenario standards could be established for *healthcare providers* in a similar manner to that of *website hosting or cloud computing providers*, which have “availability guarantees” (95%, 99.8%, 99.9%, 99.99%, etc), at increasingly higher monetary costs.

Giving more autonomy to nursing personnel has been cited as one of the possible strategies for healthcare cost reduction [103], especially in low severity cases that usually can be managed with fast-track strategies and lean planning [104]. Since ED visits in Spanish hospitals are predominantly low severity [20], some patients could be treated exclusively by nursing personnel, reserving medical interventions for high-severity cases, which would result in additional cost savings. Then models such as the ones presented in this chapter would be of even more use.

2.5.1. Limitations

Special care should be paid to personnel allocation for high-severity cases (such as traffic accidents with severe multiple trauma injuries or cerebrovascular accidents). For most of these conditions, no reliable forecast can be generated and the consequences of understaffing can be very high. Hence, a minimum level of specialized nursing personnel for severe conditions should be present at the ED (for example: 1 nurse specialized in cerebrovascular accidents, available *at all times*). On the other hand, the actual need for such personnel in the ED depends on the structure, resources and protocols followed in each hospital setting. Patients with certain clinical conditions should be hospitalized immediately. However, if this cannot be guaranteed, provisions must be made to provide adequate care for these patients in the Emergency Department.

No model can predict all events that could cause sudden increases of ED visits (such as terrorist attacks or accidents with many casualties). Hence, contingency procedures to absorb large unforeseen shocks must always exist, such as the possibility of employing on-call personnel or temporarily displacing personnel from hospitalization wards to the ED.

2.6. Conclusions

In the case of *maximum* ED census forecasts, M5P was the best choice for the reduction of major and medium nursing personnel understaffing risks, for all forecast horizons. Compared to the *usual staffing levels*, personnel planning with M5P could reduce major understaffing (>9 nurse) risks more than 10-fold (a reduction to ~1% with M5P compared to ~13% obtained with the usual nurse staffing levels); and could reduce medium understaffing (7-9 nurses) risks approximately 3-fold (a reduction to ~3% with M5P compared to ~10% obtained with the usual staffing levels). The usage of M5P would have also implied 5% - 6.1% increases in estimated nursing personnel costs (compared to the usual staffing levels), which are acceptable given the large understaffing risk reductions. This fulfilled our objective of significant reduction of understaffing with a less than 10% cost increase.

In the case of *average* ED census forecasts, once again, M5P was the best choice for the reduction of major and medium underestimation risks, for all forecast horizons. Relative risk reductions were similar to those of maximum ED census forecasts (more than 10-fold reduction in major underestimation and approximately 3-fold reduction in medium underestimation; both compared to the usual resource allocation policies). However, in this case, the usual resource allocation policies already had low risks of major and medium underestimation. On the other hand, in this case, M5P led to a *cost reduction* of more than 15%, compared to the usual resource allocation policies. Thus, the objective of reducing major and medium underestimation risks was fulfilled since the differences between M5P and usual resource allocation policies were statistically significant as confidence intervals did not overlap, for all forecast horizons; even though the absolute differences were not large.

We believe that our results prove that it is possible to produce a decision support system for the allocation of ED nursing personnel based on time series models reliant on the M5P algorithm. Forecasts can be produced with sufficient advance. We have tested a variety of forecast horizons (1 week, 2 weeks, 4 weeks and 8 weeks).

3. Prediction of inpatient admission from the emergency department (ED)

3.1. Introduction

3.1.1. The case for decision support systems in Medicine

The performance of human experts under different situational conditions has been a relevant subject of research in many disciplines. Some of the most cited findings are those of Kahneman and Tversky [105], who studied extensively the topic of human judgment under uncertainty. Their research suggests that human experts are often subjective and affected by systematic biases, and that objective decision criteria (such as scoring methods) can outperform human experts [106] in many situations.

Many areas of Medicine are fruitful application fields for decision support systems based on machine learning algorithms. For instance, there is some evidence that radiologists interpreting mammograms do not improve over time [107]. In fact, their sensitivity (in the sense of “true positive rate”) may even decrease over time, especially if they do not routinely read mammograms [107, 108]. Interpreting mammograms is challenging since the classification problem has a large degree of uncertainty in some cases [109, 110], unknown to most patients. Artificial vision systems are still in their infancy and are not ready for fully automated clinical decision making, however, in limited conditions, they have shown equivalent or even higher sensitivity and specificity than those of experienced radiologists [111, 112].

Automated decision support systems are likely to exhibit much higher concordance, lower bias and lower variance than human experts, since the former are not subject to factors that may decrease human performance such as fatigue, emotional distress or boredom [113]. However, automated decision support systems are limited by the data they may rely on and their operating conditions. Predictive models used as data-driven decision support systems usually need sufficiently large datasets to produce reliable predictions. Most importantly, they usually cannot react well to unexpected situations. However, if the problem is well-defined and the risks incurred in unexpected situations are acceptable, decision support systems based on reliable data with extensive validation can often outperform human experts [106, 114].

The prediction of inpatient admission (hospitalization) from the emergency department right after triage (and before being seen by a physician) is a problem with a sizable degree of uncertainty performed by human operators, who are frequently under stress [21-23]. This problem was deemed to be an adequate application domain for classification algorithms based on machine learning techniques, based on expert knowledge and previous research [115-117].

3.1.2. Inpatient admissions from the ED and ED triage systems

One of the main problems of hospital management is the variability of unforeseen hospital bed demand and its impact on scheduled healthcare activities [20]. For instance, unplanned inpatient admissions from the ED may result in cancellations or adjournments of surgeries that require hospitalization. Slow admissions from the ED may also result in ED overcrowding, accumulations of boarded patients in the ED, increased pressure on clinical personnel and increased clinical risks [12]. The knowledge of probabilities of inpatient admission (hospitalization) for *patients of all age groups and all triage chief complaints* present in the ED at a given moment could be helpful for streamlining of ED operations such as fast-track inpatient admissions, outbound transfer planning and rapid discharge from the ED. These predictions could also serve as an early warning system for clinical personnel in the ED and hospitalization wards.

In Spain, patients visiting EDs of large specialized care hospitals provide administrative information at arrival, which is registered and validated by administrative personnel (this may be called the administrative *check-in*). Then, patients are triaged by a specialized nurse (**Figure 17**) and wait to be seen by a clinician, with critical cases receiving medical care immediately [20] (and sometimes bypassing triage). This is the typical workflow for EDs in specialized care hospitals in developed countries. The information obtained during the administrative check-in and at triage can be of use for admission prediction. However, previous research has shown that admission prediction from the ED is a task that human experts cannot perform reliably using their clinical judgment alone [21-23].

Triage is an important component of ED workflows. Its origins are rooted in battlefield medicine. The first mention of the process is attributed to Baron Dominique Jean Larrey [118], a prominent military surgeon of the Napoleonic era. The original French term *trier*

implied sorting the wounded into three categories: those who were likely to live regardless of the surgeon's intervention, those likely to die regardless of the surgeon's intervention and those whose lives could be saved if they received timely medical care. This crude concept evolved through many years of civilian and military medical practice, resulting in sophisticated systems capable of detecting early warning signals of serious clinical conditions and stratifying patients in 3 to 5 categories of urgency of care [119]. From the point of view of algorithmic taxonomy, all major triage systems currently in use are rule-based classifiers, designed by committees of physicians and nurses with an extensive knowledge of emergency medicine. Most triage systems rely on a combination of *chief complaints* (main reasons of medical encounter) and measurements of vital signs. For instance, apparent disorientation (the chief complaint in this case), a heart rate of > 90 beats per minute, combined with a carbon dioxide tension of < 32 mmHg and either low ($< 36^{\circ}\text{C}$) or high ($> 38^{\circ}\text{C}$) body temperature may be indicative of *sepsis* [120], a whole-body inflammatory process of an infectious nature, which should be assigned a high priority of medical attention. Armed with this knowledge, triage designers may include verifications of heart rate, carbon dioxide tension and body temperature measurements in the triage decision algorithm.

It is important to note that the goal of ED triage systems is to identify patients who require the most urgent care (those who can wait *less* to be seen), not necessarily patients with the most severe underlying clinical conditions [119] or patients likely to require an inpatient admission. An example of discrepancy between the two may be a patient with an eyeball injury with a small metal shard that requires removal, with no other chief complaints or clinical conditions. The urgency of care in this situation is high (high triage priority), however, overall clinical severity is low, and hospitalization (inpatient admission) is highly unlikely since care can be delivered in an ambulatory setting.

3.1.3. The Manchester Triage System (MTS)

Before modern triage systems became available, triage assessments were performed based on clinical intuition, rather than a structured methodology. This created two major problems: the results of the triage classification were not reproducible and could not be audited [121].

The triage system currently in use in most specialized care public hospitals that possess an emergency department in the Madrid Health Authority (“*Servicio Madrileño de Salud*”) is the Manchester Triage System (MTS). The MTS is widely adopted in Spain [122], Germany, Switzerland, Austria, Holland, Ireland, Australia [121], and, to a lesser degree, in Norway and Italy [123]. It is the national triage system in the United Kingdom and Portugal [123]. It also has some presence in Brazil, Chile and Mexico [123]. Several large international studies have addressed its predictive capacity for inpatient admissions and mortality [124-126]. Other triage systems are also in use in other Spanish Regions (“*comunidades autónomas*”); however, very often, these have not been subject to large international validation studies or even small national validation studies. A recent survey shows that the MTS is the most widely used triage system in Spain *with ample international validation* [122]. We believe that the development of an admission prediction system based on the MTS can significantly ease the validation and usage in actual environments of the models developed in this work, as opposed to other less validated triage systems.

In the MTS, levels of priority establish the minimum time to be seen by a physician (**Table 14**). In practice, these times cannot always be honored due to a variety of factors, such as the number of clinical personnel actually available in an Emergency Department at the time of triage and the physical organization of the hospital. However, at the very least, these levels establish clear priorities of care.

Triage level	Name	Color	Maximum time
Level 1	Immediate	Red	0 minutes
Level 2	Very urgent	Orange	10 minutes
Level 3	Urgent	Yellow	60 minutes
Level 4	Standard	Green	120 minutes
Level 5	Non urgent	Blue	240 minutes

Table 14. Manchester triage system's priority levels.

Although triage system could provide an initial diagnosis, the MTS has focused on *priority of care* given that the variables required for an accurate diagnosis usually cannot be obtained in a short triage encounter and that a chief complaint may not be a precise description of the actual symptoms (for instance the chief complaint “abdominal pain” can be associated with many different diagnoses) [121].

The classification system of the MTS works in the following way: the triage operator selects a chief complaint and then chooses from a limited number of signs and symptoms associated to that chief complaint. Signs and symptoms that change clinical priorities are called discriminators and are presented in the MTS manuals in the form of flowcharts. Those that may affect higher priorities are sought first [121].

The MTS manual also mentions that differences exist between *clinical* priority and *managerial* priority, since some patients (such as inebriated patients, frequent visitors, social visits and others) may require special workflows *despite* their triage score [121].

3.1.4. The prediction of inpatient admission immediately after triage

Although most ED departments in Spain possess a triage system [122], to the best of our knowledge, no predictive models have been developed, that could be used in automated decision support systems for the estimation of probabilities of inpatient admission (hospitalization) for *patients of all age groups and all triage chief complaints, right after triage* (**Figure 17**), reliant on variables which were easy to obtain in any ED setting. Although some descriptive and predictive models have been published [24, 127], these usually relied on variables which are not easy to obtain in the hospital setting in immediately after triage (or are not available at all). Most of the previous models also exclude large groups of patients such as pediatric [26] or non-geriatric [24] patients. Our goal was to develop a predictive system for *patients of all age groups and all Manchester Triage System chief complaints* based on variables that could be realistically obtained in any emergency department setting for every single visit, right after triage. Some of the previous models have used post-triage variables available after a clinical evaluation performed by a physician, such as the patient location prior to ED discharge (“emergency room”, “pre-admission area”, “consultations area”, “observation area”, etc.), which are very informative for the inpatient admission outcome [128], but are useless in a real-time decision support system since, in most cases, when these variables become available, the clinical decision to hospitalize or not has been already taken. Our goal was to develop a system that could aid decision making in real-time, right after triage, before patients are seen by a physician (and thus with considerable clinical uncertainty).

Figure 17. Inpatient admission prediction *immediately after triage*.

When designing such decision support systems, attention also must be paid to contingency planning. Backup systems and procedures should always exist in EDs [129]. Logistic regression nomograms are representation of predictive logistic regression models, which show variable importance at-a-glance and can be used to calculate the probability of event manually (please see **Chapter 4** for more details). We decided to build a logistic regression model for this problem and then also a nomogram that could be used for predicting admission probabilities manually, as a contingency procedure, in case of failure of an information system supporting more sophisticated decision support systems based on machine learning models.

Besides the organizational and data availability considerations mentioned so far, several additional requirements were formulated for this classification problem:

- (i) The objective was to obtain an acceptable classifier with limited information. Near-perfect results in terms of discrimination were not attainable; since the clinical information used to make the decision to admit was not available right after triage.
- (ii) Patients who bypassed triage altogether were usually high severity cases that were almost guaranteed inpatient admissions. These were excluded from the models in order to avoid artificially inflating their performance.
- (iii) Due to the large size of the dataset, algorithms capable of processing large datasets in a reasonable amount of time were necessary. Models that were fast to train were preferable, since these could be re-generated as part of contingency procedures.
- (iv) Algorithms that did not require complex parameterization were preferable since this eased adaptation to different ED settings.
- (v) It was desirable to use models that possessed an implementation that could be deployed in diverse software environments and could be easily adapted to different information system interfaces and architectures (by means of communication through data files, web services and direct embedding in existing applications).
- (vi) Realistic error evaluation. Many techniques such as K-fold cross validation or dataset partitions in train, validation and test sets by means of uniform random sampling assume that the dataset is reasonably similar, and that the system subject to modeling is time-invariant. This may not be the case in all emergency

departments. Some circumstances that affect admission rates and change over time are hard or impossible to capture reliably with independent variables. For example: some experienced staff may be unexpectedly missing, there may be some changes in clinical protocols, chronically ill patients may stop following their treatment regimes due to worsening economic conditions, there may be intentional or unintentional bed blocking in the hospitalization wards, as well as many other situations. Thus, if partitions of the data based on uniform random sampling are used, train, validation and test datasets are more likely to be similar [130]. Therefore, *ordered* partitions of the dataset are preferred when splitting into train, validation and test sets. However, random partitions may be used within *train* sets for model derivation purposes.

3.2. State of the art

3.2.1. Calibration versus discrimination. Hosmer-Lemeshow

“groups of risk”

Machine learning and data mining model performance is usually evaluated with discrimination metrics (area under ROC, specificity, sensitivity, etc.), while the accurate estimation of predicted probabilities (or goodness-of-fit), also known as calibration, is rarely taken into account in machine learning solutions to classification problems [131-136]. However, for most binary decision-making problems, calibration is critical [137]. We may cite credit default risk or risk of cancer recidivism as examples. If an algorithm was developed to assess the risk of binary outcomes given a set of input variables, we would usually prefer a prediction of “0.51” to have a different implication in decision making than a prediction of “0.97”, although, if the cut-off point was set at “0.50” we would classify both of the aforementioned cases as *positive*. Of course, if discrimination is near-perfect, i.e. the area under the receiver operating characteristic (AUROC) is very close to 1, there is no need for further analysis. However, in real-life predictive problems, such classifiers are seldom possible, since there is usually at least some information that is not captured by any combination of independent variables. In these cases, it is helpful to possess an output which represents the probability of event accurately.

Initially, one may think that an algorithm with a good discrimination (large AUROC) is also guaranteed to have good calibration. A counter-example is a binary classifier that produces only two output values (examples: 0 and 1, 0.5 and 0.55, 0.1 and 0.2), with excellent (but not perfect) discrimination [138]. Since the ROC curve can be constructed by moving the decision threshold and calculating the resulting true positive rate (TPR) and false positive rate (FPR) at each point, such a classifier’s ROC curve would be similar to the one

presented in **Figure 18**. Since there would be only two possible output values, these would not be particularly informative when making a decision based on this classifier in *dubious* cases. In other words, this classifier’s calibration would not be adequate if the problem required an estimation of the probability of event.

Figure 18. ROC curve of a binary classifier with only two output values.

It is important to note that a classifier with good calibration also exhibits adequate discrimination – if goodness-of-fit is adequate, it is easy to set a decision threshold. However, a trade-off has been shown to exist in some cases: when nearing very high discrimination ($\text{AUROC} \gg 0.9$), calibration is likely to decrease [139]. On the other hand, in the more common case of problems with a sizable degree of uncertainty, where near-perfect discrimination cannot be achieved, calibration can be a very useful complement to discrimination. In most practical scenarios, models with both high discrimination and calibration are attainable.

Ideally, calibration metrics should compare predicted probabilities with real underlying probabilities. As the latter are usually unknown, the most common approach to assess calibration is to compare frequencies of observed (actual) and expected (predicted) outcomes. The Hosmer-Lemeshow (H-L) “groups of risk” approach [140] is a well-known measure of calibration. It is commonly used for binary problems in logistic regression

models [141], but can be used with any model that produces an output that can be interpreted as a probability. A table of observed and expected frequencies is calculated separating them by g groups. Then, for a binary classifier, the C statistic can be estimated as [140, 142]:

$$C = \sum_{k=1}^g \left[\frac{(o_k - e_k)^2}{e_k(1 - e_k / N_k)} \right] \sim \chi_{g-2}^2$$

where

g = number of groups

e_k = expected number of cases in group $k = \sum_{i=1}^G p_i$ = sum of *predicted probabilities* in group k

o_k = observed number of cases in group $k = \sum_{i=1}^G e_i$ = sum of *actual events* in group k (1 if event; 0 otherwise)

N_k = number of observations in group k

Using an extensive set of simulations with logistic regression models, it was proved that, C is usually well approximated as a chi-square (χ^2) distribution with $g - 2$ *degrees of freedom* [143]. Given these results, the H-L test, based on a p-value, was defined as follows [140]:

$$\text{p-value} = 1 - P(\chi_{g-2}^2 \leq C)$$

In this case, a p-value of *more* than 0.05 implies a lack of evidence of significant differences between the distributions of observed and expected frequencies. “ C values” are alternatively referred to as “H-L chi-square scores/values” or “H-L chi-square” or simply “H-L χ^2 ”.

Notice that: (i) the number of *degrees of freedom* changes with the number of probability groups (g) that the model is capable of producing; (ii) given the structure of the chi-square (χ^2) distribution, for fixed values of C , if the *degrees of freedom* are increased $P(\chi_{g-2}^2 \leq C)$ will decrease; (iii) for larger samples, it will be easier to find differences between the expected

and observed probability distributions (C is likely to be higher) and thus, for larger samples, it will be very hard to achieve p-values associated with very good calibration (p-values > 0.05).

Groups may be formed as percentiles of predicted probabilities (such as 1st decile, 2nd decile, ..., 10th decile) or as fixed values of the predicted probabilities (for example: 0.1, 0.2, ..., 1.0) [140, 141]. The second approach is more challenging, since it is more likely to yield null expected frequencies in some groups ($e_k=0$ for some k), and is thus more prone to produce non-computable results. We used the second approach, with fixed probability intervals, since (i) fixed probability intervals had been used in some of the previous research for similar problems [25]; (ii) for this particular problem it was not reasonable to assume that the actual probability distribution would have zeroes in any of the ten fixed probability intervals; (iii) it was easier to compare calibration *between* different models with calibration estimations based on fixed probability intervals than with calibration estimations based on predicted probability percentiles.

The H-L “groups of risk” p-value interpretation has been criticized since it may both overestimate and underestimate calibration [144, 145]. If the number of subjects in one or more groups is near to zero, H-L χ^2 may increase to infinity. In small samples, near-perfect or perfect coincidences in certain groups may occur leading to overly optimistic results – an overestimation of calibration performance. For large samples, *small* differences between groups, with no practical significance, may lead to large H-L χ^2 values. In such cases, the H-L p-value approach may result in an underestimation of calibration performance [145]. Therefore, in large samples, the $p > 0.05$ for the H-L “groups of risk” criterion may be relaxed [146], using instead tabulations of expected and observed instances by groups (calibration tables) or their graphical representations (calibration graphs), as well as raw H-L χ^2 values.

The effect of overall sample size (and thus the size of groups N_k) on H-L χ^2 values can be better understood with an example. We can see a calibration table for a *fictitious* classifier tested on a large dataset (**Table 15 A**). The calculations yield H-L $\chi^2 \approx 294.74$, which seems to imply rather poor calibration. We then divide N_k , e_k , and o_k by 10, simulating a similar classifier tested on a smaller sample (**Table 15 B**). Now H-L $\chi^2 \approx 29.18$. Repeating the same division once again, (**Table 15 C**) a smaller H-L $\chi^2 \approx 3.41$ is obtained, with a H-L p-value > 0.05 .

We can now represent the calibration tables (**A-C**) graphically, with calibration graphs: we shall represent e_k / N_k (“predicted proportion”) on the X axis and o_k / N_k (“observed proportion”) on the Y axis. A perfect calibration would place all dots on the 45° line. Given this definition, it’s easy to see that the calibration graphs for all models (**Table 15 A-C**) would be almost identical (**Figure 19 A**). Sometimes, the size of the circles on the calibration graphs is used to represent N_k (**Figure 19 B**). We thus see that H-L χ^2 values should be considered in the context of sample size. To the best of our knowledge, no “sample size normalization” criteria exist for H-L χ^2 . One possible measurement could be an “H-L R^2 ”, wherein an R^2 was calculated on the points of the calibration graph (**Figure 19 A**). If all points were on the perfect calibration line and H-L R^2 would be 1. In this example, for all cases (**Table 15 A-C**), H-L $R^2 \approx 0.92$ - 0.93 . However, this measurement has no formal justification and is only reasonable if all groups can be formed. If $e_k=0$, this would imply that $o_k = 0$, and thus “H-L R^2 ” could be high due to perfect coincidences of zeros, when calibration was actually inadequate. Since in our case the sample size was large, we used the minimization of H-L χ^2 values as a criterion of good calibration.

Even though H-L metrics has several limitations, there are few alternative calibration measurements applicable to generic classifiers [139]. In fact, viable alternatives usually concentrate on providing algorithm-specific solutions [147].

Finally, it is also relevant to mention an important empirical result. The number of groups (g) does affect the H-L χ^2 , however, with *fixed* probability intervals, the differences between H-L χ^2 with a higher g and a lower g are smaller than intuitively expected. For example, let us envision a classifier that could distribute the data similarly to the one of (**Table 15 A**), but was only capable of doing so with five predicted probabilities: 0.1, 0.3, 0.5, 0.7 and 0.9 (thus forming only 5 probability groups). Let us suppose that it is possible to collapse the predicted probabilities of **Table 15 A** into five groups (**Table 16**). We can then see that H-L χ^2 are actually quite similar for both cases: H-L $\chi^2 \approx 294.74$ with 10 groups (**Table 15 A**); and H-L $\chi^2 \approx 280.87$ with 5 groups (**Table 16**).

Group	N_k	e_k	o_k	$\frac{(o_k - e_k)^2}{e_k(1 - e_k / N_k)}$
[0 , 0.1)	9,509	963.42	597	155,07
[0.1, 0.2)	8,332	1185.19	843	115,18
[0.2, 0.3)	5,956	1355.44	1,297	3,26
[0.3, 0.4)	6,463	2057.73	2,000	2,38
[0.4, 0.5)	4,305	1722.46	1,823	9,78
[0.5, 0.6)	3,868	1984.18	1,995	0,12
[0.6, 0.7)	1,662	1008.58	1,048	3,92
[0.7, 0.8)	524	395.68	394	0,03
[0.8, 0.9)	978	803.35	786	2,10
[0.9, 1.0]	941	891.93	897	2,90
H-L $\chi^2 \approx 294.74$ H-L $R^2 \approx 0.93$				

(A)

Group	N_k	e_k	o_k	$\frac{(o_k - e_k)^2}{e_k(1 - e_k / N_k)}$
[0 , 0.1)	951	96.34	60	15,25
[0.1, 0.2)	833	118.52	84	11,72
[0.2, 0.3)	596	135.54	130	0,29
[0.3, 0.4)	646	205.77	200	0,24
[0.4, 0.5)	431	172.25	182	0,92
[0.5, 0.6)	387	198.42	200	0,03
[0.6, 0.7)	166	100.86	105	0,43
[0.7, 0.8)	52	39.57	39	0,03
[0.8, 0.9)	98	80.34	79	0,12
[0.9, 1.0]	94	89.19	90	0,14
H-L $\chi^2 \approx 29.18$ H-L $R^2 \approx 0.93$				

(B)

Group	N_k	e_k	o_k	$\frac{(o_k - e_k)^2}{e_k(1 - e_k / N_k)}$
[0 , 0.1)	95	9.63	6	1,53
[0.1, 0.2)	83	11.85	8	1,46
[0.2, 0.3)	60	13.55	13	0,03
[0.3, 0.4)	65	20.58	20	0,02
[0.4, 0.5)	43	17.22	18	0,06
[0.5, 0.6)	39	19.84	20	0,00
[0.6, 0.7)	17	10.09	11	0,20
[0.7, 0.8)	5	3.96	4	0,00
[0.8, 0.9)	10	8.03	8	0,00
[0.9, 1.0]	9	8.90	9	0,10
H-L $\chi^2 \approx 3.41$ H-L $R^2 \approx 0.92$				

(C)

Table 15. The effect of sample size on H-L calibration tables and H-L χ^2 values based on *fixed* probability intervals. (A) Large sample size; (B) Medium sample size; (C) Small sample size.

(A)

(B)

Figure 19. Examples of calibration graphs.
(A) Circles of equal size; **(B)** Circle sizes used to represent N_k .

Group	N_k	e_k	o_k	$\frac{(o_k - e_k)^2}{e_k (1 - e_k / N_k)}$
[0 , 0.2)	17,841	2148.606317	1,440	265.69
[0.2, 0.4)	12,419	3413.171786	3,297	5.45
[0.4, 0.6)	8,173	3706.646205	3,818	6.12
[0.6, 0.8)	2,186	1404.255477	1,442	2.84
[0.8, 1.0]	1,919	1695.281	1,683	0.76
H-L $\chi^2 \approx 280.87$ H-L $R^2 \approx 0.92$				

Table 16. The effect of the number of groups (g) on H-L calibration tables and H-L χ^2 values based on fixed probability intervals.

3.2.2. Algorithms capable of producing adequately calibrated models

Classification algorithms capable of producing models with adequate *discrimination* do not necessarily produce models with adequate *calibration*. We shall discuss some algorithms that produce predictions that can be interpreted as probability distributions and thus are more likely to produce adequately calibrated classifiers: logistic regression, LogitBoost and Artificial Neural Network with cross-entropy cost function and softmax output layer.

3.2.2.1. Logistic regression

Logistic regression optimizes log likelihood and bounds the output $p(x)$ to the $[0, 1]$ range:

$$p(x) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 x_1 + \dots + \alpha_N x_N)}} = \frac{1}{1 + e^{-(LP)}}$$

where

α_0 = constant term

N = number of independent variables

$\alpha_{i=1..N}$ = independent variable coefficients

$x_{i=1..N}$ = independent variable values

LP = “linear predictor”

If event probabilities can be approximated with a sigmoid function given a set of predictors, which is very common in many predictive problems in Medicine, this approach can produce calibrated classifiers.

The main assumption for the usage of a logistic regression is that the *logit* of the outcome variable is a linear combination of the independent variables. Multicollinearity must be

avoided and there should be sufficient *events* of the dependent variable for its combinations with independent variables (“cells”). Otherwise, numerical optimization algorithms may not converge. Also, similarly to many other regression procedures, outliers can affect the model in a significant manner.

It is possible to prove that a logistic regression problem can be solved as a *convex minimization* problem [148]. *Convex minimization* problems have several desirable properties, but the most relevant are: (i) if a *local* minimum exists, it is a *global* minimum and (ii) if there is a minimum in a *strictly convex function*, that minimum is unique. Numerical methods used for the calculation of logistic regression coefficients can make use of these properties for reaching a global optimum solution in a fast manner. However, in the case of logistic regressions, non-convex problems may arise, especially with highly correlated variables or small datasets with many of variables.

Logistic regression implementations in MATLAB and Weka are similar. The Stata implementation is more flexible, having many execution options that Weka and MATLAB lack, such as the possibility of specifying diverse numerical optimization methods or computing *robust* confidence intervals. It also has adjustments for ill-defined models (such as models with insufficient events in certain “cells”) and is also significantly faster than the Weka implementation. For well-defined problems, without multicollinearity and with a sufficient number of events in each cell, all three software packages should produce almost identical independent variable coefficients.

3.2.2.2. LogitBoost

LogitBoost is an algorithm for the optimization of *additive* logistic regression. The additive logistic regression has the following form for two-class problems [149]:

$$\log \frac{P(y = 1 / x)}{P(y = -1 / x)} = \sum_{m=1}^M f_m(x) = F(x)$$

where

$f_m(x)$ = simple function (*weak learner*), such as a *decision stump*.

M = number of $f_m(x)$. A new $f_m(x)$ is generated within each iteration (**Algorithm 1**), and is also referred to as “number of iterations” (as in the case of the Weka implementation [150]).

This expression can be transformed into a form similar to the logistic regression:

$$p(x) = P(y = 1 / x) = \frac{e^{F(x)}}{1 + e^{F(x)}} = \frac{1}{1 + e^{-F(x)}}$$

The algorithm for a two-class problem functions as follows [149]:

1. Start with weights $w_i = 1 / N$ $i=1, 2, \dots, N$, $F(x) = 0$ and probability estimates $p(x_i) = 1/2$
2. Repeat for $m = 1, 2, \dots, M$:
 - (a) Compute the working response (z_i) and weights (w_i):

$$z_i = \frac{y_i^* - p(x_i)}{p(x_i)(1 - p(x_i))}$$

$$w_i = p(x_i)(1 - p(x_i))$$
 - (b) Fit the function $f_m(x)$ by a weighted least-squares regression of z_i to x_i using weights w_i .
 - (c) Update $F(x) \leftarrow F(x) + \frac{1}{2} f_m(x)$ and $p(x) = \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}}$
3. Output the classifier $\text{sign}[F(x)] = \text{sign}[\sum_{m=1}^M f_m(x)]$

Algorithm 1. Logit Boost Algorithm.

The algorithm has two parameters: the number of iterations (M) and the type of function used for $f_m(x)$. Usually, if the number of iterations is increased, from a certain point, there is no further gain in classification performance metrics, but there is also no worsening. Hence, a high value of the number of iterations is almost always a safe bet. As for the other parameter $f_m(x)$, *weak learners* such as decision stumps (single-split trees with only two terminal nodes) function very well in practice and are often the best choice [149].

LogitBoost uses Newton steps for fitting the aforementioned expression, similarly to a logistic regression, by maximum likelihood [149].

It has been shown empirically that LogitBoost *with decision stumps* as weak learners usually exhibits a high resistance to overfitting, performing well even with correlated predictors [149, 151, 152]. The authors of the original implementation of LogitBoost state that a fully formalized theoretical justification of its resistance to overfitting is still lacking,

but provide several hypothetical explanations [149]: (i) with more LogitBoost iterations, the overall impact introduced by each $f_m(x)$ is reduced, since only observations with appreciable weight determine the new functions; (ii) the nature of the boosting algorithm (with decision stumps as weak learners) does not allow for the full collection of variables to be *jointly* fit, which results in lower variance.

LogitBoost has been used for many different applications such as prediction of fraud in mobile advertising [153], magnetic resonance image analysis [154] or text genre classification [155].

3.2.2.3. Artificial Neural Network with cross-entropy cost function and softmax output layer

The *inspiration* for Artificial Neural Networks (ANNs) is the high-level behavior of neurons in the human brain, giving the possibility of modeling complex non-linear relationships between variables [156]. Structurally, ANNs consist of processing units interconnected with weighted connections. In the case of feed-forward ANNs, the high-level structure is that of an input layer, wherein independent variables are introduced; an output layer, from which the dependent variables are obtained; and one or more hidden layers that serve as a connection between them. One of the most commonly used methods for training ANNs in an efficient manner is the backpropagation algorithm. It works as follows. In the process of training, weights are assigned to connect inputs and outputs. Then, output values are compared to the correct answer and the incurred error is computed. The error information is then sent back through the network and used to adjust the weights in such a way that this error is reduced. After a number of iterations, the aforementioned error is typically reduced to a small amount [156]. Depending on the implementation, several parameters, including number of hidden layers, number of neurons in each hidden layer, learning rate may be used to adjust ANNs' structure, calculation time as well as discrimination and calibration capacities for a given problem [157].

The main shortcomings of ANNs are the difficulty in interpretability and the difficulty of finding optimal solutions due to local minima in most ANN structures. Indeed, a significant amount of research has been devoted to designing heuristics [158] to avoid local minima in the training process.

ANNs have been used with a large variety of classification problems, such as the prediction of diabetes [159, 160], breast cancer classification [161, 162], credit scoring [163], word-

confidence estimations in speech recognition systems [164], among many other applications [165, 166].

The first published mention of an ANN was that of McCulloch and Pitt [167]. They used a threshold neuron as a model for information storage and processing in the human brain.

Figure 20. McCulloch and Pitts neuron

In this approach, all inputs were combined into a single number Z .

$$Z = b + \sum_{i=1}^N w_i x_i$$

where

b = bias term

x_i = attribute i

w_i = weight associated to the attribute x_i

N = number of attributes

McCulloch and Pitt defined a threshold neuron that performed a simple prediction in the following manner:

$Y = 1$ if $Z > 0$,

$Y = 0$ if $Z \leq 0$.

This function was not differentiable, which posed difficulties for training the neuron [168]. Perceptrons [169] were a further development which replaced the threshold neuron by a continuous function $g(Z) = Y$, called *activation function*. This function is usually chosen in such a way that it is differentiable, thus easing the training process [168].

Figure 21. Perceptron

Many different activation functions are possible. Some of the most common are: linear, sigmoid, hyperbolic-tangent and arc-tangent:

The linear function is $g(Z) = Z$

The sigmoid function is $g(Z) = \frac{1}{1 + e^{-Z}}$

The hyperbolic-tangent function is $g(Z) = \tanh(Z) = \frac{2}{1 + e^{-2Z}} - 1$

The arc-tangent function is $g(Z) = \frac{2}{\pi} \arctan\left(\frac{\pi \cdot Z}{2}\right)$

The hyperbolic-tangent function and the sigmoid are linearly related. In fact, the hyperbolic-tangent function is also called *hyperbolic-tangent sigmoid* (or *tan-sigmoid*) function. Several authors argue that the hyperbolic-tangent function should be used instead of the sigmoid activation function, since it has desirable properties for ANN training [170].

Some authors also showed empirically that ANN training with the hyperbolic-tangent function was faster and that results were better on certain datasets [171].

Multilayer feed-forward ANNs are usually formed by interconnected perceptrons.

A relevant question is how to produce outputs bounded to 0 and 1 for multiclass outputs. In the binary case, a logistic output may be used, since then this output can then be interpreted as the probability that the input data belong to one of the two choices. For two or more classes, a more general approach is a *softmax* activation function (also called *softmax layer*) [168]:

$$y_i = \frac{e^{Z_i}}{\sum_{j=1}^C e^{Z_j}}$$

where C is the number of categories of the output variable.

The softmax activation function has the following properties:

- (1) $\sum_{j=1}^C y_j = 1$
- (2) $0 < y_i < 1$

Thus, softmax activation functions can produce outputs that can be interpreted as probability distributions across discrete alternatives [166].

Notice that the *softmax* activation function is not just normalization, since it involves the exponentiation of Z .

Another relevant topic is the correct choice of an *error function* (also called *cost function*) for ANNs. The chosen error function should be adequate for the purpose of the ANN and should have a short training time. The usage of sum-of-squared error function is based on the assumption of a continuous output and Gaussian noise. Thus, its usage with binary outcomes is not recommended, since the output is not continuous and errors usually follow a Bernoulli distribution [168].

Modeling the binary classes as Bernoulli observations led to the use of the cross-entropy [172, 173]. For the binary scenario, the cross-entropy error function is [168]:

$$E = -\sum_{i=1}^N [t_i \cdot \ln(\hat{t}_i) + (1 - t_i) \ln(1 - \hat{t}_i)]$$

where

N = number of samples

t_i = observed output (1 if “yes”; 0 if “no”)

$\hat{t}_i$ = predicted output (1 if “yes”; 0 if “no”)

3.2.3. Previous research

The prediction of inpatient admission from the ED has been addressed in previous research.

The most representative approaches we have found are shown in **Table 17**:

- (i) A non-MTS admission prediction model [25] was developed in Singapore. The reported AUROC was 0.849 (0.847-0.851 95% CI) on the validation dataset and 0.850 (0.848-0.852 95% CI) on the derivation (train) dataset and an H-L $p > 0.05$. However, the partition between the validation and derivation datasets was based on uniform random sampling, which could result in an optimistic estimation of model performance [130]. Also, this model was based on a regional *ED patient acuity classification* (PAC) and variables reflecting chronic conditions which may not be routinely registered as part of the triage process in other countries.
- (ii) Another non-MTS based model was developed in Spain for predicting inpatient admission and readmission only for *elderly patients* (≥ 64 years of age), with a point estimate (without confidence intervals) of the AUROC of 0.76 on the test set.
- (iii) A large-sample prediction model for admission from the ED, reliant on variables available right at the end of the MTS triage process, was recently developed in the United Kingdom [26]. A 6-variable scoring method was derived from a logistic regression model. This scoring method produced an AUROC of 0.8774, (0.8752 - 0.8796 95% CI) on the validation set and an H-L p -value > 0.05 for the comparison of admission rates between train and test sets on more than 35 “score levels”. However, this model was restricted to *adult patients* (≥ 16 years of age). In addition, the usage of the National Health Service National Early Warning Level, which is exclusive to the United Kingdom, and the varying interpretation of triage categories in the different hospitals in this study hinder its applicability in other countries.

Two additional models are *not* shown in **Table 17**:

- (iv) An MTS-based study was performed in the Madrid Region (Spain) in 2013 [128]. This study had several limitations: it had a relatively small sample (2,476 cases) and *did not* use a test set for the estimations of predictive performance, i.e. performance metrics were reported *only for the train set*. Its applicability was also restricted by the *exclusion* several major clinical specialties and all ED visits from 10 PM to 8 AM.
- (v) Finally, a recent paper described a model for 30-day inpatient readmission prediction, developed on a large database of historical inpatient admissions in the state of California, U.S.A. [174], by means of CART with logistic regressions on its leaves (the authors use the term “conditional logistic regression”, which is more common in Statistics). The objective of this work was *different*, and results were measured only in terms of prediction accuracy = (true positive + true negative) / total, with an improvement of ~20% over a standard logistic regression model. However, even with these differences, the algorithm by itself was deemed to be an interesting approach from an algorithmic point of view.

	Year	Country	Target population	Train / test dataset sizes	Independent variables	Performance on the test set	Notes
[24]	2011	Spain	Elderly patients (≥ 65 years)	<p>Total size of the population = 28,430 individuals with ED visits in 2006, 2007 and 2008</p> <p>Unreported number of visits (1 visit per individual assumed; criteria of individual choice unclear).</p> <p>Unreported train and test set sizes.</p> <p>Random partition of data for the generation of train and test sets.</p>	<ul style="list-style-type: none"> - Sex - Age: 4 categories - Insulin-dependent diabetes - Chronic obstructive pulmonary disease - Ischemic heart disease - Heart failure - Number of concurrent diagnoses - Number of drugs prescribed - Emergency admissions 2007 - Planned admissions 2007 - Days of cumulative stay 2007 - Emergency admissions 2006 - Planned admissions 2006 - Days of cumulative stay 2006 	<p>AUROC = 0.76 (point estimate, no CI)</p>	<p>Logistic regression model.</p> <p>Calibration reported as H-L p-value = 0.50 with unreported number of groups (unclear if this calculation was performed on the train or the test set; most likely this is the value obtained on the train set since SPSS was used).</p> <p>Data from two hospitals (Costa de Ponent healthcare area, Catalonia region, Spain).</p>
[25]	2011	Singapore	Patients of all age groups	<p>Train = 190,549 visits (60%)</p> <p>Test = 127,032 visits (40%)</p> <p>Random partition of data for the generation of train and test sets.</p>	<ul style="list-style-type: none"> - Age group - Race - Arrival mode: Walk-in, ambulance - PAC: Singapore triage score - ED visits in preceding 3 months - Hospital admission in preceding 3 months - Chronic conditions: 7 categories 	<p>AUROC = 0.849 (0.847-0.851 95% CI)</p> <p>H-L p-value > 0.05 (with 10 fixed probability groups)</p>	<p>Logistic regression model.</p> <p>Data from one hospital (Singapore).</p>
[26]	2014	United Kingdom	Adult patients (≥ 16 years)	<p>Train = 215,231 visits (66.66%)</p> <p>Test = 107,615 visits (33.33%)</p> <p>Random partition of data for the generation of train and test sets.</p>	<ul style="list-style-type: none"> - Age group: 9 categories - NEWS (NHS National Early Warning Score): 11 categories - Triage category: 1, 2, 3, 3+ - Referred by GP: yes / no - Arrived in ambulance: yes / no - Admitted < 1 year ago: yes / no 	<p>AUROC = 0.8774 (0.8752 - 0.8796 95% CI)</p> <p>H-L p-value > 0.05 (with 35 "score levels")</p>	<p>Initially, a logistic regression model with more variables is generated. Then, a scoring system is derived, with just 6 of the initially considered variables.</p> <p>The performance reported on the test set is that of the scoring system.</p> <p>Data from three hospitals (North Glasgow region, United Kingdom).</p>

Table 17. Similar previous models for the prediction of inpatient admission from the Emergency Department.

3.3. Materials and Methods

3.3.1. Data source and study setting

Predictive models were developed within a retrospective observational study performed in the Ramón y Cajal University Hospital of Madrid (Spain), a specialized care referral center with all medical specialties excepting Obstetrics. In years 2011 and 2012 it had approximately 1,100 beds and 553,000 patients assigned to its clinical area (one of the largest healthcare areas of the Madrid Region). The research project was authorized by the hospital's Ethics Committee for Clinical Research and by the Deputy Managing Director. Research protocols were established in accordance to national legislation and jurisprudence for Clinical Research and Patient Rights. The variables used in the predictive models were routinely collected by administrative personnel and triage nurses.

A custom data management system was constructed in QlikView to extract, combine, transform and validate the information required for this study. Administrative check-in data for patient arrivals to the ED in years 2011 and 2012 were obtained from the HP-HIS central hospital information system. Triage level and chief complaints for the same period were obtained from the HNS-HCE Emergency Department information system. Then, prior surgeries and ED visits were obtained for each patient and index ED visit. Patient gender and age at the time of the ED were obtained from the latest extract of the master patient registry. Visit source and visit cause variables were normalized following the criteria of the Emergency Care Minimum Dataset (MDS) of Emergency Care of the Madrid's Health Service. Arrivals by ambulance were detected with a text parser used on a commentary field of the ED visit table. Patient insurance status was obtained following the Madrid Health Service's inter-hospital billing guidelines. Each ED episode was joined with the corresponding monthly snapshot of the healthcare insurance card database of the whole

Madrid Region (~6.500.000 registers within each snapshot), using the healthcare insurance card number and “year-month” combination as the join field. This provided an accurate estimation of patient insurance status *at the moment of the ED visit*. Healthcare insurance card numbers were obtained from weekly snapshots of the master patient registry of the Ramón y Cajal University Hospital (~600.000 registers within each snapshot). Patients’ regions of residence were also obtained from these snapshots.

Patients without a valid triage score, as well as patient discharges where the medical decision to admit could not be observed were excluded (patients dead on arrival, outbound patient transfers, patients leaving without being seen and voluntary discharges). The few visits with chief complaint “Pregnancy” were also excluded, since the sample was deemed too small and non-representative given the lack of a maternity ward in the Ramón y Cajal University Hospital. After the exclusions, the dataset contained a total of 255,668 ED visits (~94.57% of all ED visits).

The infrequent *exitus letalis* (deaths) that occurred in the emergency department were considered inpatient admissions, since these cases should have resulted in an inpatient admission, but did not for a variety of causes (rapid death, lack of vacant beds in hospitalization wards, etc.).

The dataset was ordered by date and time of arrival to the ED and then split, with the first 66% of the dataset used for the derivation of the predictive models (168,740 ED visits), the next 14% of the dataset for validation (35,793 ED visits) and the remaining 20% for testing (51,135 ED visits). In previous research, partitions of data into train and test sets were often based on uniform random sampling [24-26], but these may be more likely to produce optimistic results since train and test sets will be generally more similar than with an ordered split [130].

3.3.2. Optimization objective

We had addressed this same problem (inpatient admission prediction from the ED) in a previous publication [175]. We had used 9 independent variables and developed two models: a logistic regression model and an ANN model with a cross-entropy cost function and a softmax output layer. In that case, 66% of the data were used as a train set and the remaining 34% were used as a test set. The ANN model was derived with *default* parameters: 10 neurons in a single hidden layer, with no regularization. The results we had obtained on the test set were AUROCs of 0.8568 (95% CI 0.8508 - 0.8583) for the logistic regression model and an AUROC of 0.8575 (95% CI 0.8540 - 0.8610) for the ANN model.

H-L χ^2 values on the test set for Hosmer-Lemeshow (with 10 fixed probability groups, or, conversely, 8 degrees of freedom for the H-L χ^2 distribution) were 65.32 for the LR model and 17.28 for the ANN model. A statistically significant difference between AUROCs for the ANN and LR models could not be established. However, calibration was better for the ANN model.

We wanted to further increase the AUROC, compared to our own model [175]. We also wanted to outperform a similar model, for all patients and all chief complaints, developed in Singapore (mentioned in **Section 3.2.3**), that had also achieved a result of approximately 0.85 [25]. Thus, the discrimination objective for this problem was defined as $\text{AUROC} > 0.85$.

At the same time, we wanted to reduce H-L χ^2 as much as possible (with 10 fixed probability groups, or, conversely, 8 degrees of freedom for the H-L χ^2 distribution). As stated previously, for very large samples, such as this one, the H-L p-value > 0.05 criterion was very hard to achieve, since even extremely small differences could lead to p-values < 0.05

(Section 3.2.2). Given 10 fixed probability groups, i.e. 8 degrees of freedom, this was equivalent to H-L $\chi^2 < 15.5073$.

The need of forming 10 fixed probability groups was important for the practical usage of this predictive system since different users could employ different decision boundaries on the Probability of inpatient admission (P_{adm}). For instance, bed reservations could be made if $P_{adm} \geq 0.70$. If $P_{adm} \geq 0.80$ and the patient could not be hospitalized for lack of available beds in the hospital, outbound patient transfer planning could be started. Elderly patients could be sent to an observation area if $P_{adm} \geq 0.50$. On the other hand, patients could be considered for “rapid discharge” from the ED if $P_{adm} \leq 0.10$. These criteria may vary, but there is an obvious need for adequate calibration with a sufficiently large number of divisions in the predicted probabilities.

We wanted to achieve improvements over previous models by means of: (i) including more informative variables; (ii) improving the parameterization of algorithms; (iii) considering other existing algorithms; (iv) considering ad hoc algorithms for this problem.

3.3.3. Independent variables included *a priori*

Independent variable selection (also called features selection) is an unsolved problem in Statistics and Machine Learning. If domain knowledge is available, it should be used as the first step of the analysis in order to select relevant predictors and to design ad hoc features [176].

Dimensionality reduction techniques, such as principal component analysis (PCA), are often used in classification problems with a large number of possibly redundant features that have no clear interpretation. However, in our case, we decided *not* to use dimensionality reduction techniques since: (i) we wanted to use few variables with a plausible interpretation (so that the system could also be used manually, as a contingency procedure); (ii) we had expert knowledge of the problem and could also rely on variable selections performed in previous research.

We decided to use *categorization* in all variables. We used some of the variables selected in previous research (**Section 3.2.3**) and analyzed risk levels of the variables on sub-partitions of the train set in order to define categories with reasonably uniform levels of risk of admission.

The first variables we considered were those of the Manchester Triage System (MTS): triage levels and triage chief complaints.

Since there are 50 chief complaints in the MTS (2nd edition) [121], we decided to group these (excepting chief complaint “Pregnancy”) into a more manageable number of categories. Chief complaint definitions [121] and raw admission rates on the *train* dataset were be examined. Then, chief complaints were combined in five groups of increasing risk of

admission. Cut-offs between groups were defined at $\leq 2.5\%$, $\leq 5\%$, $\leq 10\%$, $\leq 40\%$ admission percentages (**Table 18**).

Broad visit causes (traffic accident, occupational accident, poisoning, aggression and common disease) may also serve as a proxy of risk of admission. Also, patient visit sources (referral from primary care, patient decision, hospital department referral, referral from another hospital) and ambulance arrivals were included since they are known to affect risks of admission [25, 26].

Patient insurance status and main residence were included in the models in a variable with three categories: social security insurance and main residence in the hospital's region; social security insurance in patients from other regions and a third category which combined all other financing options. Our hypothesis was that differences could exist between these groups (for instance, patients with social security insurance from other regions could exhibit higher severity than patients from the hospital's region).

Group	Chief complaint	ED visits	% Inpatient admissions
Excluded	Pregnancy		
S0 ≤2.5%	Sexually acquired infection	72	0.00%
	Burns and scalds	110	0.91%
	Eye problems	12.011	0.96%
	Ear problems	3.226	1.21%
	Neck pain	2.453	1.22%
	Rashes	4.102	1.83%
	Assault	919	2.18%
	Torso injury	1.072	2.24%
S1 [2.5 % - 5%)	Foreign body	230	2.61%
	Back pain	6.410	2.82%
	Nasal problems	1.271	3.15%
	Dental problems	1.840	3.42%
	Apparently drunk	365	3.56%
	Exposure to chemicals	53	3.77%
	Irritable child	130	3.85%
	Bites and stings	167	4.19%
	Headache	2.539	4.25%
	Sore throat	2.237	4.38%
	PV bleeding	1.186	4.55%
	Testicular pain	539	4.64%
	Worried parent	1.170	5.04%
	Wounds	3.034	5.31%
S2 [5% - 10%)	Limb problems	33.060	5.36%
	Unwell child	5.120	5.43%
	Abdominal pain in children	1.392	5.60%
	Limping child	82	6.10%
	Head injury	2.296	6.27%
	Crying baby	100	7.00%
	Overdose and poisoning	525	8.57%
	Urinary problems	6.067	8.64%
	Falls	3.130	8.82%
	Major trauma	995	10.15%
	Shortness of breath in children	1.110	10.54%
	Abscesses and local infections	3.809	10.61%
	Diarrhoea	1.432	13.34%
	Self-harm	202	17.33%
S3 [10% - 40%)	Abdominal pain in adults	12.681	17.85%
	Behaving strangely	990	17.88%
	Asthma	136	18.38%
	Collapsed adult	2.828	18.81%
	Vomiting	1.719	18.85%
	Mental illness	768	20.44%
	Fits	522	21.84%
	Chest pain	6.818	25.14%
	Unwell adult	25.651	25.48%
	GI bleeding	1.124	26.25%
	Haematological illness	173	26.59%
	Diabetes	575	36.52%
S4 >40%	Shortness of breath in adults	10.299	51.10%

Table 18. Classification of chief complaints by risk of admission on the train set.

Repeated visits within 365 days prior to the index visit were included with categories 0-1 visits [24], 2-12 visits and 13 or more visits. The last category can be described as that of very frequent users [177] (sometimes called *frequent flyers*), who may be less likely to be admitted in their frequent and often clinically unjustified visits. These may include worried parents of newborns, socially excluded persons or certain types of elderly patients.

Surgeries may result in side effects resulting in an urgent inpatient admission. Since some surgical treatments may result in complications over time, we decided to include surgeries within the last 4 months prior to the ED visit as a variable. More than two surgeries in this time period are infrequent; hence we decided to include a variable with three categories: 0 surgeries, 1 surgery and 2 or more surgeries within 4 months. Although it may seem that non-elective (urgent) surgeries could result in a higher risk of inpatient admission from the ED, this is not the case if risks of complications are similar for both elective and non-elective surgeries. Non-elective surgeries may increase pre-surgical length of stay, since they are not scheduled in advance (hence some physical and human resources may thus not be initially available), but they do not necessarily increase the risk of inpatient admission. We thus decided to include both elective and non-elective surgeries.

Some studies have shown an increase of adverse clinical outcomes in unscheduled patient care dependent on weekdays [178, 179] or different time shifts [180]. These results have been subject to debate and controversy, but the fact remains that, emergency departments (and hospitalization wards) in all developed countries are almost always unequally staffed during different shifts and weekdays. This could also have an effect on admissions. We thus created a variable with two categories: “weekend or night shift” = Y / N.

It is important to use variables that can be observed in combination with the outcome in training data subsets. If a categorical variable’s level is not observed in the training process,

the numerical method of the algorithm may have convergence problems. In fact, some software packages simply omit the variable (or the variable *level* in the case of categorical variables) altogether in this situation. This situation may occur if a categorical variable with a large number of categories, some of them infrequent, is included in the dataset (such as the 50 chief complaints of the MTS 2nd edition) or if variable interactions are used, such as an interaction between normalized chief complaints with triage score.

3.3.4. Independent variables excluded *a priori*

We had initially considered the inclusion of several additional independent variables, but finally decided not to include them due to various considerations. We believe that mentioning the criteria followed for the exclusion of these variables *a priori* is useful both for understanding our approach and for future research.

Previous inpatient admissions and previous bed days within a certain time interval are not necessarily a good predictor for posterior urgent admissions, since they lack information about clinical severity. Indeed, some *clinical pictures* (all the information relating to a disease, disorder, or a patient's state) require hospitalization, but are not severe and are not likely to result in a subsequent non elective admission. Hence, some clinical severity adjustments could be made, but they require coding of the episode in a structured language, such as the International Classification of Diseases (ICD-9, ICD-10 and ICD-10-PCS). The number of plausible combinations between ICD codes is enormous. Diagnosis-Related Groups (DRG) were developed in order to combine these possibilities into a manageable number of categories (usually between 500-2000 categories for all possible clinical cases). An initial version was created in the Yale University in the early 1970s, called Yale DRGs. From then on, many different DRG systems have been developed in different countries. Unlike France, Germany or England, Spain did not develop a national DRG system, relying instead on DRG systems developed in the United States [181]. Medicare DRGs (MS-DRGs and CMS-DRGs); All Patients DRGs (AP-GRDs); All Patients Refined DRGs (APR-DRGs) [182] have been used in Spain for managerial purposes. To the best of our knowledge, until now, International Refined DRGs (IR-DRGs) [183] have only been used for research projects in Spain.

Initially, we wanted to include variables based on APR-DRGs and IR-GRDs, since both DRG systems possess risks of mortality levels, and higher risks of mortality in previous hospitalization episodes are likely to be correlated with higher risks of admission in subsequent contacts with healthcare services. However, risks of mortality are not normalized between DRGs. For instance, a risk of mortality of 2 in the “APR-DRG 316 – Hand & Wrist Procedures” is not equivalent to a risk of mortality of 2 in the “APR-DRG 001 – Liver transplant”. Risk of mortality normalizations can be performed. However, DRG risk of mortality databases are not always available in a timely manner, especially for the version of DRG grouper in force at a given moment.

DRGs in general also present other methodological and practical issues: (i) their primary objective is to classify by *average prospective costs of care*, while risk of mortality assessments are secondary; (ii) adherence to standard coding practices varies, resulting in large differences in coding quality between hospitals [184] and even between individual coders within the same hospital; (iii) as coding languages evolve, so do DRGs. Models reliant on DRGs would need to account for these changes and have to be re-calculated with the latest version of the grouper. While differences between grouper *versions* (within the same *family*) can be significant, differences between grouper *families* can be dramatic (for instance the differences between AP-DRG and APR-DRG families [6] or AP-DRG and IR-DRG families [183]). Sometimes, DRG families are simply abandoned and a new DRG family is chosen instead, with no possibility of using the previous DRG system with subsequent updates in coding languages. For instance, in 2016, national and regional healthcare authorities in Spain decided that the AP-GRD family was to be completely replaced by the APR-DRG family, after more than 20 years of using AP-DRGs [185]. All these considerations must be taken into account when designing models based on DRGs.

Charlson and Elixhauser comorbidity indexes adapted to ICD coding [186, 187], as well as other clinical variables based on ICD coding, may be used. However, ICD coding is subject to a relevant practical problem - in Spain, and in most European countries, coding is done after the patient has been discharged and after the discharge report has been written and validated. Unless coding is done by the discharging physician, there is a certain backlog (of days or weeks, and sometimes *months*) between patient discharge and coding. Although semi-automated ICD coding systems have been proposed [188-193], no realistic alternative to human coders exists yet for a general scenario, especially with such large and complex dictionaries as those of ICD-10 and ICD-10-PCS. Fully-automated translation from clinical reports to ICD codes is very challenging due to ambiguities, semi-technicalities, abbreviations, orthographical omissions and localisms in clinical reports. Also, the risk of incorrect ICD coding can be high for many applications. Therefore, human ICD coders are likely to be needed in the foreseeable future and, unless coding is done immediately after patient discharge, decision support systems that require the latest clinical information (such as emergency care decision support systems) should *not* rely on variables based on ICD codes and DRGs.

3.3.5. Feature space exploration

Feature selection methods in the field of Machine Learning are classified in filter methods (algorithm-independent) and wrapper methods (algorithm-dependent) [194]. Hybrid methods also exist. We chose to perform feature selection for the logistic regression and use the resulting variable selection for all other algorithms. Several methods for feature selection exist for regression models. Stepwise regression is perhaps one of the most popular ones, with three approaches: (i) forward selection, (ii) backward elimination and (iii) bidirectional elimination. Usually, forward selection is recommended if nothing is known about the dataset. These methods have several problems [195-197], with some of authors even stating that “stepwise” regression should be called “unwise” regression [198]. LASSO (Least Absolute Shrinkage and Selection Operator) is an alternative technique that performs both regularization and variable selection [197]. However, the most general method of variable selection for a given model is to try every possible combination of independent variables to select the best solution for the target optimization metric (this approach is sometimes called “all possible regressions procedure”). Of course, this approach is not practical if the computational time required is prohibitive. Fortunately, logistic regression implementations usually have fast convergences on large datasets in most modern machine learning and statistical software packages (such as Weka, Stata or MATLAB).

Several implementations of the “all possible regressions” procedure exist, although one of the fastest and most flexible we have found is the **allsets** command, implemented for Stata [199]. It allows testing different variable combinations for linear, logistic and Cox regression models. There were two shortcomings in the original implementation: (i) it used an implementation of H-L with deciles, instead of fixed groups; (ii) it only allowed calculating performance metrics on the *train* dataset. We modified the **allsets** command so that it did

perform a calculation H-L χ^2 values based on fixed probability intervals and performed this calculation of metrics for *validation* datasets.

We included the 11 independent variables previously mentioned: triage score, triage chief complaint group, age range, gender, surgeries in the past 4 months, weekend or night shift, visit number in the last 12 months, insurance status, visit source, visit cause, ambulance arrival. The triage score was set as a *fixed* variable, to be included in all models. A total of 1,024 models were generated as a result (all possible 11-variable combinations with triage score set as a fixed variable, which are 1,023 plus one model with just the triage score). The following metrics, for train and validation datasets were stored for each model: Akaike information criterion (AIC), Bayesian information criterion (BIC), Brier score, AUROC, H-L χ^2 and degrees of freedom of the H-L χ^2 distribution. The performances of the models were plotted as a function of AUROC and H-L χ^2 on the train set (**Figure 22 A**) and on the validation set (**Figure 22 B**).

In terms of calibration, models which did not achieve at least one sample in each one of the 10 fixed probability groups (and thus 8 degrees of freedom) were excluded from the analysis. On the *train* set this occurred in 108 models, all of them with less than 7 variables. On the *validation* set, this occurred in 135 models.

The models on the *validation* dataset (**Figure 22 B**) had a lower variability than those of the train set. However, the best models on the *train* set generally performed worse on the *validation* set (**Figure 22 A**). Also, models that failed to produce 10 groups (or, conversely, 8 degrees of freedom of the H-L χ^2 distribution), were omitted from both figures and there were *more* of these models on the *validation* set than on the *train* set (135 versus 108 models).

Finally, the effect of the sample size must be taken into account. Since the size of the train set (168,740 ED visits) was larger than the size of the validation set (35,793 ED visits), some reduction in calibration could also be due to these differences.

The highest AUROC with the lowest H-L χ^2 was obtained for the model that included all 11 independent variables on the train set. The AUROC of the 11-variable model on the *validation* dataset (**Figure 22 B**) was comparable to that of the *train* set. The H-L χ^2 of the 11-variable model on the *validation* set was slightly better (AUROC ≈ 0.862 and H-L $\chi^2 \approx 40$) than that of the *train* set. This seemed to suggest that the 11-variable logistic regression model was reasonably stable and we chose to use all 11 variables, under the assumption that it would be possible to further improve calibration (i.e. decreasing H-L χ^2) while maintaining or even improving discrimination (i.e. increasing AUROC) with other algorithms. We used the 11 variables in all subsequent experiments.

(A)

(B)

Figure 22. Comparison of model performance on the (A) train and (B) validation datasets. *108 models with less than 8 degrees of freedom are excluded on the train dataset (A) and 135 models are excluded on the validation dataset (B).*

3.3.6. ANN and LogitBoost parameterization

In the ANN and LogitBoost parameterization experiments, the first 66% of the dataset were used for the derivation of the predictive models (168,740 ED visits) and the next 14% of the dataset for validation (35,793 ED visits). As stated in **Section 3.3.5**, all 11 variables were used. The ANN and LogitBoost algorithms were chosen due to their capability of producing adequately calibrated classifiers (**Section 3.2.2**).

3.3.6.1. ANN parameterization

The ANN model was developed in MATLAB, with the **patternnet** command - a feedforward network with a hidden layer, which uses the hyperbolic tangent sigmoid activation function on the neurons of its hidden layer, and a *softmax* activation function [200] on its output layer. Its outputs can be interpreted as probability estimates. This ANN implementation possesses certain resistance to overfitting. This is achieved as follows. The *training* data, i.e. the train set, fed to **patternnet** is split in two sub-divisions: 70% of the data for a subtrain set and 30% of the data for a subvalidation set. In our particular case, the train set (168,740 ED visits) was split into two subdivisions generating a 70% subtrain set (118,118 ED visits) and a 30% subvalidation set (50,622 ED visits). In **patternnet**, the train subdivision is used for updating the network's weights and biases. The errors on the train and validation subdivisions are monitored through the iterations. If the error on the validation subdivision begins to increase, overfitting is assumed. Network weights and biases are saved at the minimum of the error observed on the validation subdivision, thus avoiding overfitting in most cases [200]. The partition of the dataset may be random, ordered or interleaved. An ordered partition implies using the last samples of the train set for internal validation, i.e. not actually using these data for training, but for internal adjustment. In this particular problem, this did not seem to be a valid approach, since this would have implied throwing

away a large percentage of recent and potentially useful information. An interleaved division did not seem to have a theoretical justification in our case either. In fact, it could result in the omission of some repeating phenomena, such as visits during certain time periods. This left the default option of random partitioning between subdivisions. Notice that a random partition of the dataset for *training* purposes is legitimate (as mentioned in **Section 3.1.4**). On the other hand, classification metric performance on validation or test sets should be evaluated on ordered partitions of the data.

patternnet has a very fast implementation that allows hardware parallelization with CPUs *and* GPUs, which makes it several orders of magnitude faster (on modern computers) than non-parallelized ANN implementations, such as Weka’s Multilayer Perceptron.

patternnet has only two parameters: number of neurons in the hidden layer (which we shall call “hidden neurons”) and a regularization parameter. The range of possible values of the regularization parameter is [0-1], with 0 meaning no regularization. The minimum number of hidden neurons is 0. Independence between these two parameters could not be assumed. Therefore, we performed joint testing for both parameters. We tested the interval between 0 and 50 hidden neurons, and regularization between 0 and 1, with 40 divisions.

Classification performance metrics were obtained for the full train (168,740 ED visits) and validation datasets (35,793 ED visits), as previously described in **Section 3.3.1**. Results were displayed as graphics using a scattered interpolation procedure for 3-dimentional graphics implemented in MATLAB [201].

Discrimination was very similar on the train and validation sets (**Figure 24**). However, there were differences in H-L χ^2 between train (**Figure 23 A**) and validation sets (**Figure 23 B**) and it was hard to establish stable optimums (low values of H-L χ^2) that held on both

sets. The white areas in **Figure 23 A** and **Figure 23 B** are models that did not achieve 8 degrees of freedom (i.e. could not form the 10 probability groups).

Although a previous **patternnet** model we had developed, with less independent variables (9 independent variables), exhibited adequate calibration with default parameters (hidden neurons = 10; regularization = 0, i.e. no regularization) [175], in this case, calibration was inadequate with default parameters both on the train and validation sets (**Figure 23**). We chose hidden neurons=46, regularization=0.1, since this seemed to be a possible “stable valley” in terms of calibration. For this parameterization, the following metrics were obtained on the validation dataset: AUROC = 0.869; H-L $\chi^2 \approx 20$

(A)

(B)

Figure 23. Comparison of *patternnet* ANN H-L χ^2 on the (A) train and (B) validation datasets. 0-50 hidden neurons and 0-1 regularization. *Only models with 8 degrees of freedom are shown.*

(A)

(B)

Figure 24. Comparison of patternnet ANN AUROC on the (A) train and (B) validation datasets. 0-50 hidden neurons and 0-1 regularization.

3.3.6.2. LogitBoost parametrization

The next algorithm we wanted to try was LogitBoost [149] since it is usually resistant to overfitting, and is generally capable of outperforming a logistic regression (LR) or, at least, matching its performance.

Boosting algorithms usually perform well with the adequate kind of weak learners. The resistance to overfitting of LogitBoost is believed to be associated to weak learners such as *decision stumps*, that do not perform *joint* optimization of variables, as this is thought to result in higher possibilities of overfitting [149]. We did not find suitable alternatives to decision stumps in this particular case.

Experiments on the train set showed little improvement with more than 150 iterations of LogitBoost with decision stumps as weak learners (**Figure 25**). It was observed that H-L χ^2 were lower on the *validation* dataset than on the *train* set. Part of the decrease in H-L χ^2 could be attributed to the smaller size of the validation set.

On the validation set, results stabilized with the number of iterations on AUROC ≈ 0.862 and H-L $\chi^2 \approx 58$, i.e. no improvement over the logistic regression model with all variables.

Figure 25. Evolution of LogitBoost performance on train and validation datasets
(A) AUROC; (B) H-L χ^2 from 0 to 400 iterations; (C) H-L χ^2 from 50 and 200 iterations.

3.3.7. Ad hoc classifier derivation

We also decided to develop an ad hoc classifier for this problem. As with the previous experiments, the first 66% of the dataset were used for the derivation of the predictive models (168,740 ED visits) and the next 14% of the dataset for validation (35,793 ED visits). Also, once again, all 11 variables were used.

3.3.7.1. SimpleCART tree with logistic regressions on its leaves

After the experiments with LR, ANN and LogitBoost, we decided that, perhaps, data could be separated with CART trees and different algorithms could be subsequently applied on its leaves (similarly to an approach mentioned in **Section 3.2.3**). Initially, we used the SimpleCART [202] classifier in Weka to derive a series of trees. SimpleCART's advantage over J48 (a C4.5 implementation in Weka) is that it can produce splits based on several levels of a categorical variable simultaneously. For instance, given variable A with levels A1, A2, A3, A4; and variable B with levels B1, B2, SimpleCART can produce a split such as "split 1: (A=A1 or A2)". On the other hand, the J48 algorithm, can only produce splits using all variable levels (i.e. split 1: A=A1, split2: A=A2, split 3: A=A3, etc.). This usually results in SimpleCART producing smaller trees, although training is slower. Following the ideas of [174], we tested submodels with trees of 2 levels, 3 levels and 4 levels of depth. Within each final leave, we used our modification of the **allsets** procedure (as stated in **Section 3.3.5**, **allsets** was a Stata implementation of the "all possible regressions" procedure), to calculate all possible logistic regressions to obtain the optimal variable selection within each leave. Unfortunately, the results we obtained in all cases were *worse* than those of a simple 11-variable logistic regression model on the *train* set (with AUROC<0.75 and poor calibration).

3.3.7.2. Ensemble of ordinary logistic regression and MTS chief complaint tree with LogitBoost on its leaves

Since the purely data-driven approach to data splitting did not produce improvements over the 11-variable logistic regression model, we decided to split data based on triage *chief complaints*. We deemed this approach reasonable since: (i) the Manchester Triage System decision-making process was organized by chief complaint and (ii) some of the independent variables we were using could have different importance in the context of different chief complaints. For instance, possibly, previous surgical episodes did not have the same relevance in the case of the “G.I. (gastro-intestinal) bleeding” chief complaint and in the case of the “Rashes” chief complaint. Similarly, an ambulance arrival in the case of the “Abdominal pain” chief complaint, most likely, did not have the same importance as in the case of the “Fits” chief complaint.

We then decided to use logistic regression (LR) submodels on chief-complaint leaves. These LR submodels (one logistic regression for each chief complaints submodel) often failed to converge, and the number of iterations of the numerical optimization algorithm (in the case of Stata, the default is the Newton-Raphson method with step halving and special fixes when non-concave regions of the likelihood are encountered [203]) had to be artificially set to 15 to stop optimization in non-converging models. This was likely due to some chief complaint submodels not providing sufficiently large datasets for convergent models with 10 independent variables (11 variables minus the chief complaint itself, which would not be included in the submodel). However, even with these restrictions, discrimination and calibration metrics improved on the *train* set, with a point estimate of AUROC = 0.8802, and H-L = 12.50. Discrimination and calibration on the *validation* set were AUROC = 0.8663, and H-L = 71.36, with a calibration that was *worse* than a simple LR with all variables on the *validation* dataset (**Figure 22**). Such large differences could be indicative

of overfitting. Also, the usage of a certain number of iterations for *artificial* stopping of optimization in logistic regression classifiers was not deemed to be an elegant solution.

There could be several ways of enhancing the LR submodels. One was the usage of stepwise procedures with backward selection. However, this approach was problematic since between-term collinearity was guaranteed to occur in at least some chief complaints, and most stepwise procedures would stop when this happened, aside from the theoretical problems associated with stepwise procedures in general [195-198]. Another option was the usage of **allsets** in all leaves, thus examining all possible variable combinations in each chief complaint leaf. However, we decided not to pursue this path since it was computationally cumbersome (49 chief complaints x 512 possible models on each leaf) and could lead to overfitting. The other alternative was the usage of LogitBoost *with decision stumps*, since this approach is generally highly resistant to overfitting and can usually match or outperform a LR based on the same data [149]. As stated before, LogitBoost was likely to produce adequately calibrated models. Previous results with LogitBoost for the whole dataset implied that optimal results were achieved with 150 or more iterations (**Section 3.3.6.2**). Preferring to err on the side of safety, we executed LogitBoost submodels with 200 iterations.

However, for certain chief complaints, a LR model with all 11 variables (we shall call it the “base LR model”) could be better than the LogitBoost submodel. In order to account for this eventuality, we decided to compare the resulting models and tried to choose the best between them (**Algorithm 2**). First of all, we calculated the base LR model with the whole *train* dataset (M1). Then, we calculated LogitBoost submodels for each chief complaint (M2_{cc}) on the same *train* dataset. We then optimized for calibration directly. For each chief complaint submodel, we calculated the H-L χ^2 values and degrees of freedom of the base LR (model M1) and the H-L χ^2 values and degrees of freedom produced by the chief-complaint

submodel developed with LogitBoost (model M2_{CC}). Then, the model with the highest number of degrees of freedom and the lowest H-L χ^2 values was chosen. Recall that degrees of freedom are calculated subtracting 2 from the number of groups that could be achieved with the predicted values (**Section 3.2.1**). Given that we were using fixed probability intervals, we deemed reasonable to consider M2_{CC} over M1 only if the number of groups (and thus degrees of freedom) achieved with M2_{CC} was higher. The comparison of raw H-L χ^2 between distributions which may have different numbers of groups is an empirical solution (**Section 3.2.1**; **Table 15 A** versus **Table 16**). The full algorithm can be formally described in pseudo code as follows:

```

1. Compute M1 = base logistic regression for the whole dataset
2. FOR EACH CC = Chief complaint
    Compute M2CC = LogitBoost submodel for this Chief complaint
    Compute H-L DF |M2CC = H-L degrees of freedom for M2CC
    Compute H-L  $\chi^2$  |M2CC = H-L  $\chi^2$  for M2CC
    Compute H-L DF |M1 = degrees of freedom for M1 in the subset corresponding to this Chief complaint
    Compute H-L  $\chi^2$  |M1 = H-L  $\chi^2$  for M1 in the subset corresponding to this Chief complaint
    IF ( (H-L DF |M2CC >= H-L DF |M1 ) AND (H-L  $\chi^2$  |M2CC <= H-L  $\chi^2$  |M1) )
        Use M2CC for this chief complaint
    ELSE
        Use M1 for this chief complaint
    END IF
END FOR
3. Output the predictions of the ensemble

```

Algorithm 2. Ad hoc ensemble classifier algorithm for the inpatient admission prediction from the emergency department.

The point estimates of discrimination and calibration metrics were: H-L χ^2 = 16.91, AUROC = 0.8739 on the train set and H-L χ^2 = 22.77, AUROC = 0.8667 on the validation set. Discrimination was similar to that of the previous model (that may be called “chief-complaint tree of LRs with artificial optimization stopping”), but calibration was better. From now on, we will refer to the model resulting from **Algorithm 2** as “ad hoc ensemble classifier”.

We decided to use Stata for the implementation, since it offered an easier way of managing datasets and encoding variables in a consistent manner than MATLAB. Although Stata did possess an implementation of LogitBoost through a third-party plug-in (implemented in C++ for Windows) [152], it was slow even for small datasets, and thus inadequate for the dataset size requirements of this project. The Weka LogitBoost implementation was much faster, but there was no interface with Stata. In order to solve this issue, we developed a Weka-Stata interface which allowed the usage of Weka's LogitBoost from Stata. We developed a custom Java program which used the Weka API, and Stata's command line invocation and Stata's **merge** commands to join the predictions produced by Weka with the Stata datasets.

3.4. Results

Models were recalculated with the combined data of the *train* and *validation* datasets and evaluated with the data on the *test* set, using the variable categorization of triage chief complaints groups obtained on the *train* set (**Table 18**). All 11 variables were used, with the combined train and validation sets (first 80% of the dataset, 204,533 ED visits) used for the derivation of the predictive models and the next 20% for the evaluation of the results (51,135 ED visits).

3.4.1. Logistic regression odds ratios

All variables (**Table 19**) were significant in the LR model for all independent variables at $p < 0.001$. The most important predictor was the triage level, followed by the triage chief complaint group. Odds increased monotonically with the increase in triage priority and category number of the triage chief complaint group. If the ED visit originated from a different hospital or from a request of a clinical service within the same hospital, odds were higher than if the visit had originated from primary care. Patients with social security insurance were more likely to be admitted than patients with other types of insurance. Within patients with social security insurance, patients from other regions had higher odds of admission than patients from the hospital's region. Occupational and traffic accidents, as well as aggressions (associated with trauma injuries, which can be treated in an ambulatory setting) had lower odds than common disease cases. Patients with more than 13 visits in the last 365 days (frequent users) had lower odds of admission than patients who visited the ED less frequently in the same period. The number of surgeries in the last four months increased the odds of admission. ED visits during the night shift or during the weekend slightly *decreased* the odds of admission, although the overall effect was small.

	Odds ratio	95% Lower CI	95% Upper CI
MTS score			
L1	158.9458	100.1777	252.1896
L2	30.52967	19.59041	47.57738
L3	14.52033	9.327278	22.60465
L4	3.832508	2.461133	5.968032
L5	1		
MTS chief complaint (CC) group			
S0	1		
S1	2.762753	2.436437	3.132773
S2	4.344648	3.897215	4.84345
S3	8.387682	7.545357	9.32404
S4	18.49013	16.52442	20.68967
Visit during the weekend or during the night shift			
Yes	1		
No	1.143772	1.104565	1.184372
Surgeries (preceding 4 months)			
00	1		
01	1.387235	1.305583	1.473995
02+	1.937225	1.714389	2.189026
ED visits (preceding 12 months)			
13 or more visits	1		
00-02 visits	1.915889	1.578855	2.324869
03-12 visits	1.736708	1.427631	2.112701
Age range			
1-14 years	1		
0 years	2.37065	2.004197	2.804108
15-35 years	1.067634	0.971628	1.173126
36-65 years	2.006901	1.842493	2.185979
66-80 years	2.853253	2.618039	3.1096
80-90 years	3.532769	3.234471	3.858576
90+ years	4.446986	3.974978	4.975044
Gender			
Female	1		
Male	1.300235	1.261412	1.340252
Insurance status			
Other			
Social security (this region)	1.271611	1.118602	1.445548
Social security (other regions)	1.616921	1.370221	1.908037
Visit source			
Primary care	1		
From home	1.0546	0.983886	1.130398
Hospital	2.552334	2.198372	2.963288
Other hospitals	7.995346	6.551986	9.75667
Visit cause			
Traffic accident	1		
Aggression	4.44395	3.06733	6.438399
Common disease	6.847329	5.449244	8.604115
Occupational accident	2.228675	1.386517	3.582354
Poisoning	2.409939	1.508962	3.848874
Ambulance arrival			
No	1		
Yes	1.617958	1.558745	1.679421

Table 19. Odds ratios for the LR model built on the train and validation sets.

3.4.2. Calibration and discrimination on the test set

Models were re-calculated with the data from the combined training and validation datasets. Then, these models were evaluated on the test set.

Confidence intervals (CIs) for AUROCs were calculated using the methodology described in [204], by means of Stata's **rocreg** command with 1000 bootstrap repetitions in each case.

The results in terms of AUROC on the *test* set were as follows:

- The LR model produced an AUROC = 0.8534 95%CI (0.8487, 0.8582);
- The ANN model produced an AUROC = 0.8558 95%CI (0.8511, 0.8605);
- The ad hoc ensemble classifier model produced an AUROC = 0.8657 95%CI (0.8621, 0.8694)

For AUROCs, a statistically significant difference could be established between the ad hoc ensemble classifier model and the LR model since the 95% CIs for their respective AUROCs did not overlap. The estimation of AUROC of the ANN model was higher than the estimation of AUROC for the LR, however, a statistically significant difference between AUROCs could *not* be established.

The results in terms of calibration, on the *test* set were as follows:

- The LR model produced H-L $\chi^2 = 37.74$ and H-L $R^2 = 0.9953$.
- The ANN model produced H-L $\chi^2 = 36.67$; and H-L $R^2 = 0.9952$.
- The ad hoc ensemble classifier model produced H-L $\chi^2 = 14.3$, the only classifier producing an H-L p-value = 0.063 > 0.05, i.e. lack of convincing evidence of *differences* between groups (following the p-value criterion defined in **Section 3.2.1**); and H-L $R^2 = 0.9978$.

The H-L χ^2 obtained in all experiments had 8 degrees of freedom (i.e. the 10 groups of probability could be formed in all cases).

Calibration graphs were generated for all three models (**Figure 26**). The interpretation of these graphs is explained in **Section 3.2.1**. Calibration was slightly better for the ad hoc ensemble classifier model than for the LR model (**Figure 26 A** vs **Figure 26 B**). Calibration values were very similar for the LR model and the ANN model (**Figure 26 B** vs **Figure 26 C**). H-L R^2 values were similar for all three models (0.9953 for the LR, 0.9952 for the ANN, 0.9978 for the ensemble classifier).

A nomogram was generated for the LR model (**Figure 27**) using the **nomolog** command in Stata [205], which is described in more detail in **Chapter 4**.

Error distributions per chief complaint on the *test* set for the ad hoc ensemble classifier can be found in **Appendix B**. Some chief complaints exhibited higher average errors than those of other chief complaints. This was a logical result since clinical uncertainty varied between triage chief complaints in terms of general patient state and thus risks of inpatient admission.

(A)

(B)

(C)

Figure 26. Calibration on the train set.
 (A) Logistic regression; (B) ANN; (C) Ensemble classifier

Probability of inpatient admission from the E.D.

Figure 27. Nomogram representation for the logistic regression model.

3.4.3. Estimations of calibration and discrimination metrics with increasing train sets and fixed-length test sets

In order to assess the steadiness of the differences between the three algorithms (LR, ANN and the ad hoc ensemble classifier) and to simulate an approximation to real use, we decided to use an additional testing scheme. We used increasingly larger train sets with fixed-length test sets (with approximately monthly length, subsequent to the increasingly larger train sets). The first experiment used 50% of the dataset for training (approximately 1 year of historical information) and the next 1/12 of the next 50% of the data (approximately 1 month) for testing. The train set was increased by 1/12 of the remaining 50% in each experiment. Thus, 12 experiments were performed with ordered partitions of the dataset, with *test* sets that were independent from each other, in each experiment (**Figure 28**).

Both variable selection and variable categorization should be considered part of the *training* process. Thus, within each experiment, the train sets were each split at 66% into subtrain and subvalidation sets. Chief complaint categorization was repeated on the subtrain sets within each experiment, producing slightly different variations of the chief complaint severity group variable. Since we had performed a variable selection *a priori* by means of consulting prior research and expert knowledge, we decided to include all 11 variables in all experiments. Algorithm parameterization is also part of the *training* process. LR required no parameterization. The ad hoc ensemble classifier's only parameter was the number of iterations in LogitBoost submodels. It was set at 200. For each experiment, ANN was parameterized (choosing plausible values for hidden neurons and regularization) following the procedure described in **Section 3.3.6**, using subtrain sets for algorithm derivation and subvalidation datasets for parameter estimation. In each experiment, all models for all algorithms (LR, ANN and the ensemble classifier) were recalculated on the full train set of

that experiment (subtrain + subvalidation) and evaluated on the full test set of that experiment.

Performance evolution graphics of AUROC, H-L χ^2 and H-L R^2 for all algorithms were derived (**Figure 29**). Confidence intervals were calculated using the t -distribution.

Average AUROCs were:

- For LR models, average AUROC = 0.8531, 95% CI (0.8501, 0.8561),
- For ANN models, average AUROC = 0.8568, 95% CI (0.8531, 0.8606),
- For ad hoc ensemble classifier models, average AUROC = 0.8635, 95% CI (0.8605, 0.8665).

Confidence intervals for average AUROCs did not overlap in the case of the LR models and the ensemble classifier models. In the case of the LR models and the ANN models, confidence intervals for average AUROCs slightly overlapped, although the ANN models had higher AUROCs than the LR model in all but one month (**Figure 29 A**).

Average H-L χ^2 were:

- For the LR models, average H-L χ^2 = 35.15, 95% CI (32.57, 37.73)
- For ANN models, average H-L χ^2 = 10.47, 95% CI (7.78, 13.17)
- For the for ad hoc ensemble classifier models, average H-L χ^2 = 11.4, 95% CI (9.10, 13.75).

The H-L χ^2 obtained in all experiments had 8 degrees of freedom (i.e. the 10 groups of probability could be formed in all cases). Both ANN and ad hoc ensemble classifier models possessed better calibration than LR models, with H-L p-values > 0.05 in 10 of the 12 experiments.

Figure 28. Logistic regression and Ensemble classifier evaluation scheme.

(A)

(B)

**Figure 29. Comparison of Logistic regression, ANN and Ensemble classifier on an increasingly larger train set, with a 1 month test set in all cases.
(A) AUROC; (B) H-L χ^2 with 8 degrees of freedom.**

3.5. Discussion

3.5.1. Interpretation of results

The results on a test set formed with the last 20% of the initial dataset seemed to suggest that only the ad hoc classifier fulfilled the optimization objective of AUROC > 0.85 and H-L p-value > 0.05 . In this case, the 95% CIs for the AUROCs of the ANN model and LR model overlapped and their calibration metrics were very similar.

However, these results varied when considering more extensive (and more representative) experiments on increasing train sets and fixed-length test sets. In this case, the differences between ANNs and LRs were not statistically significant in terms of discrimination, but in this case, ANNs had significantly better calibration than LRs (**Figure 29**). These results were coherent with our previous research [175]. Discrimination and calibration were consistently better for ad hoc ensemble classifier models than for LR models (**Figure 29**). In this case, both ANN models and ad hoc ensemble classifier models fulfilled the optimization objectives of AUROC > 0.85 (i.e. “lower 95% CI for AUROC > 0.85 ”) and reduction of H-L χ^2 , with H-L p-values > 0.05 in 10 of the 12 experiments. The H-L χ^2 obtained in all experiments, with all classifiers (LR, ANN and ad hoc ensemble), had 8 degrees of freedom (i.e. the 10 groups of probability could be formed in all cases). As explained in **Section 3.3.2**, attaining 10 groups fixed groups of probability was very important for the practical applicability of our models.

The ad hoc ensemble classifier incorporated expert knowledge, which enabled it to achieve better results than exclusively data-driven models. It also had several desirable characteristics: ease of application, evidence of adequacy of results, scalability, safety and

protection against overfitting. It was almost self-tuning: only the number of LogitBoost iterations for chief complaint submodels had to be set to a sufficiently high value. The usage of LogitBoost in M2_{CC} (**Algorithm 2**) offered certain guarantees against overfitting. The comparison of H-L χ^2 degrees of freedom and χ^2 values obtained from M2_{CC} submodels and M1 subsets (**Algorithm 2**) provided an empirical calibration optimization algorithm.

LR models were easier to derive and to use, especially considering that they could be represented with a nomogram. However, better results in terms of both calibration and discrimination were obtained with ad hoc ensemble classifier models. In experiments with increasing train sets and fixed-length test sets, there was also an improvement of calibration in ANN models compared to LR models.

Generally, we believe that our approach exhibits adequate results for all algorithms (LR, ANN and ad hoc ensemble classifier), given that it was designed for *patients of all age groups and all chief complaints* excepting “Pregnancy”.

A direct comparison of results with previous approaches developed for other hospitals ([24-26, 127, 206], **Table 17**) was not possible due to differences in datasets, exclusion criteria and independent variables. However, to the best of our knowledge, none of the aforementioned models had achieved such high AUROCs while also maintaining adequate calibration for such an ample group of patients.

3.5.2. Possibilities of deployment of predictive system

Since non-runtime editions of Stata (Stata / IC, Stata / SE and Stata / MP) are not designed for usage as business logic containers in enterprise applications, we developed a layer to provide this functionality, bypassing the protections implemented in these editions of Stata [207, 208] to prevent this kind of usage. We had previously received an explicit permission from Stata Corp Ltd. to do this for a series of user-written commands (**nomolog**, **nomocox** [207, 208] and **coin** [209]), as long as the software wherein Stata was embedded was not used for commercial purposes. The usage of Stata-based web services was also envisioned, which could ease the introduction of this decision support system as well as general-purpose Stata applications into environments with implementations of the Service Oriented Architecture (SOA) [208].

Also, a pure Java implementation could be devised using Weka's API directly. The main advantage of a Weka-based implementation is its ease of integration in Java Enterprise Edition applications, which are increasingly common in large organizations. The advantage of using Stata instead of Weka is that it possesses implementations of many potentially useful statistical procedures that Weka lacks, as well as efficient data management functions. Also, with large datasets, an instance of Stata on a single computing node may slow down significantly, but will not crash. On the contrary, an implementation based on the standard Weka API on a single computing node can easily run out of memory with large datasets. Some solutions are the usage of systems with large amounts of Random Access Memory (RAM) and distributed computing techniques such as a distributed implementation of Weka on Hadoop.

3.5.3. Limitations

Our approach did not account for unnecessary or avoidable inpatient admissions and detecting these was not one of its goals [210].

All triage systems are reliant on the operator, as some subjectivity is always present when defining apparent symptoms and navigating through the triage algorithm's decisions [211]. However, if triage personnel is adequately trained, inter-operator concordance has been shown to be high in the case of Manchester Triage System (MTS) [212].

As emergency medicine evolves over time, so do triage systems, although most often these changes introduced are not radical and are limited to a few chief complaints. However, even if new chief complaints are introduced, this could be overcome by assigning the new chief complaint to existing chief complaint groups (**Table 18**), such as the ones that we have developed in this work.

Since the Ramón y Cajal University Hospital did not provide obstetric care during the time of the study, the triage chief complaint “Pregnancy” was excluded. However, we believe it could be assigned to chief-complaint group “S3” (**Table 18**), although it would be desirable to recalculate the models using a dataset that did include a more representative sample. Alternatively, a submodel could be derived exclusively for patients visiting the ED with chief complaint “Pregnancy”.

3.6. Conclusions

We believe that we have achieved our objective of building models suitable for an admission prediction system for all age groups and all triage chief complaints (excepting “Pregnancy”), based on few variables available right after triage. It should be possible to obtain equivalent variables in the ED of any specialized care hospital which uses the Manchester Triage System. We believe that this approach for developing decision support systems for admission prediction can also be used with other triage systems, although it would be preferable to rely on triage systems with ample scientific validation. The models developed in this work system may be used for inpatient and ED bed management, rapid discharge, fast-track admission and outbound transfer planning, as well as an early warning system for clinical personnel in the ED and hospitalization wards.

The ad hoc ensemble classifier algorithm results in models with acceptable calibration and discrimination metrics, is fast to train even on large datasets, and does not require complex parameterization – the only parameter is the number of iterations in LogitBoost leaf models, that can simply be set to a reasonably high value (such as 200). Since LogitBoost has been shown to be resistant to overfitting, we believe that the overall model, with LogitBoost on its leaves, should be resistant to overfitting too. We thus hypothesize that this approach can be considered reasonably safe and easy to use. The programming implementation of the ensemble classifier can be deployed in different IT environments. There are two possibilities of deployment: a pure Java implementation based on the Weka API or a Stata-based program with a web interface integrated with the Weka API. Logistic regression models have worse discrimination and calibration metrics than ad hoc ensemble classifier models. However, LR models can also be used as contingency procedures by means of nomogram representations [121].

4. Nomograms

4.1. Introduction

4.1.1. Nomograms

Since the late XIX century [213] many graphical calculators have been developed for a variety of engineering problems [214]. The historical name used to refer to these graphical calculators is *nomogram*. A formal definition of a nomogram can be found in [215]:

Nomograms are one of the simplest, easiest and cheapest methods of mechanical calculation. (...) precision is similar to that of a logarithmic ruler (...). Nomograms can be used for research purposes (...) sometimes leading to new scientific results.

Nomograms based on regression models are also used in clinical publications and clinical practice [216]. Multivariate logistic regressions and Cox regressions are commonly used in several scientific fields to build predictive models.

Predictive logistic regression models help to fulfill one of the goals of a regression analysis: to predict the probability of the outcome using a set of independent variables. Odds ratios tables are the most common way to present predictive models in several disciplines, but these do not allow the direct calculation of output probabilities. Full logistic regression formulas are sometimes given in article appendices, but as the number of variables in the model rises, these quickly become burdensome to apply [205].

Cox regression models are widely used in survival models in many fields, such as sociology, reliability engineering and medicine [217]. Due to its popularity, the original description of the Cox regression [218] is the one of the most cited papers in the field of Statistics. In this case, obtaining probabilities of survival for a given time is also challenging with tables of regression coefficients or with full regression formulas.

As a solution, a graphical calculator for output probability score calculation for logistic and Cox regressions was presented by Kattan [219-223]. These nomograms have become popular in several areas of clinical research (such as Radiation Oncology) and, although alternative styles of nomograms exist, *in this work* we chose to focus on Kattan-style nomograms.

In some cases, nomograms have been shown to offer higher predictive accuracy and better performance characteristics than ANNs, CART, look-up tables and risk-group stratification models [224]. Nomograms can be very useful in a busy clinical practice and can be easily introduced in clinical workflows, since they are user-friendly and can be paper-based. They also provide the possibility of including patients in a multivariate clinical decision-making process in a transparent manner (for example, a doctor may say to a patient: “given your characteristics as a patient, and our model based on averages, if you choose treatment A, which is less aggressive, your probability of survival is approximately X; if you choose treatment B, which is more aggressive, your probability of survival is approximately Y”). Nomograms should not substitute clinical judgment, but can be a useful complement to it in some decision-making contexts.

In addition to enabling users to obtain predicted values manually, nomograms provide excellent graphical depictions of all variables in the model [225], providing a quick view of the weight of each variable [205]. Compared to a black-box mechanism, wherein an operator

may introduce values of variables (for instance a spreadsheet-based calculator), a nomogram also provides a safety mechanism, since visualizing variable scales *at all times* makes it is less likely to use incorrect variable values.

As with any other statistical tool, nomograms have limitations and should not be used mindlessly. It must be said that a nomogram is a *representation* of a predictive model, not a *validation* tool. If the underlying model is not adequately calibrated (i.e. it does not exhibit an adequate *goodness-of-fit*), the output probabilities obtained with the nomogram are most likely to be meaningless, even on the *train* dataset. Nomogram users must make sure that the calibration of the model is adequate for their purposes on at least one representative validation dataset before generating a nomogram. Prior to clinical usage, calibration should be assessed on representative validation datasets.

However, even if calibration is not adequate, but discrimination is, a decision boundary can be set on the nomogram (for example, $P = 0.5$) and the nomogram can be used for discrimination purposes only (for example, decide “yes” if $P \geq 0.5$ and “no” if $P < 0.5$). Also, some scenarios could be envisioned wherein a rough separation between two decision boundaries (resulting in three decision areas, for instance “not likely”, “average”, “likely”) could also be an adequate result, although, once again, this depends greatly on its intended usage and adequacy of validation. Probability labels should be removed in the two aforementioned scenarios since these would no longer represent probabilities, but decision boundaries.

Also, nomograms should only be used for *approximate calculations*. If high-precision results are required (ex: hundredths of probabilities are relevant), alternative approaches should be used. However, such high precision requirements are not common in biomedical research, mainly, since the underlying models themselves usually cannot offer this kind of precision.

Several characteristics are desirable in nomograms: (i) even if coefficients are negative in the original regression, variables should be forced positive on the nomogram in order to avoid subtractions (for ease-of-use and error avoidance); (ii) nomograms should have the capability of supporting two-variable interactions (two-variable products), since these are sometimes used in biomedical research; (iii) event (or time-to-event) probability axes should have reasonable boundaries.

Although software packages such as the R statistical software [226] and, to a much lesser extent, SAS [227], may produce nomograms, these generators have several shortcomings and are not user-friendly. We intended to build logistic and Cox regression nomogram generators for a statistical software package that was widely used in biomedical research - Stata [228]. Our aim was to build nomogram generators entirely in the Stata and Mata [229] programming languages, without external software dependencies, implemented as a post-estimation command, to be executed after the **logistic** or **logit** commands (logistic regression commands included in Stata's core packages) in the case of logistic regression nomograms; and **stcox** commands in the case of Cox regression nomograms. This was a very laborious programming task due to the limitations of Stata's programming model and built-in graphical capabilities.

4.1.2. Logistic regression nomogram example

We may use a fictitious logistic regression model for the description of the structure and usage of binary logistic regression nomograms. Let us suppose we use variables age, sex and number of transfusions as predictors of probability of hematological complications (yes/no). The model is built using Stata's **logit** command:

logit complications gender transfusions age

This command produces the following table of *logit* coefficients and a constant term (“_cons”):

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
age	.0652398	.0069921	9.33	0.000	.0515356	.078944
transfusions	.0362445	.0115255	3.14	0.002	.0136549	.0588342
gender	.5388903	.1747807	3.08	0.002	.1963265	.8814542
_cons	-5.783012	.4558551	-12.69	0.000	-6.676472	-4.889553

Let us assume that this model has adequate calibration (recall **Section 3.2.1**) on the train set and on a series of representative test sets.

Then, the nomogram in **Figure 30** can be generated with the **nomolog** program.

The calculation of output probability using the nomogram of **Figure 30**, is then performed as follows: (i) obtain scores for all variable values; (ii) add all scores; (iii) then the probability of event is obtained using the *Total Score to Probability* conversion. For example, for a 40 year old male who had 35 transfusions, the scores for the respective variables are: $\text{Score}(\text{Male}) \approx 0.8$; $\text{Score}(35 \text{ transfusions}) \approx 1.9$; $\text{Score}(40 \text{ years old}) \approx 4$. Then, *Total Score* would be approximately 6.7, which is equivalent to a probability of event of approximately 0.22.

Figure 30. Logistic regression nomogram example.

This example illustrates several uses of logistic regression nomograms:

1) Output probability calculation with a fast and simple graphical calculation method. A nomogram, once generated, does not require a computer to obtain an estimation of event probability. This allows its use in environments and situations where computing devices are unavailable or inconvenient. A classical example is that it can enable patients to make informed treatment decisions, as they allow an easy evaluation of “what-if” scenarios [220], given that the underlying models have been validated on representative datasets and exhibit adequate calibration.

2) Variable weight is clear at-a-glance, as shorter variable scales indicate lesser relative importance. This eliminates the *black-box effect* (lack of understanding of underlying mechanisms of calculation), which is increasingly frequent in complex models. As a side-effect, the nomogram also becomes a useful tool for exploratory or descriptive analyses.

4.1.3. Derivation of logistic regression nomograms

The formula of a logistic regression can be written as:

$$p(x) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 x_1 + \dots + \alpha_N x_N)}} = \frac{1}{1 + e^{-(LP)}} = \frac{1}{1 + e^{-(\alpha_0 + TS)}}$$

where

α_0 = constant term

N = number of independent variables

$\alpha_{i=1..N}$ = independent variable coefficients

$x_{i=1..N}$ = independent variable values

LP = “linear predictor”

TS = *Total Score*

The basic derivation of the logistic regression nomogram is the following:

1. Compute all possible scores ($\alpha_i x_i$) for each variable ($x_{1..N}$).
2. Establish the displacement in the probability line (in the *Total Score to Probability* conversion) due to the constant term (α_0).
3. Transform plausible sums of scores (*Total Score*) into probabilities of event given the expression of $p(x)$.
4. Plot the nomogram.

As described in the previous section, nomograms can be used to derive probabilities of events graphically. Once variables are given specific values, the *Total Score* can be obtained. It can then be transformed into a probability of event using the lower part of the nomogram (*Total Score to Probability* conversion).

Since nomograms should be easy to use for humans several additional considerations must be taken into account: (i) individual variable scores should be rescaled to a range of 0 – 10 to ease operations, (ii) probability intervals should have reasonable boundaries, (iii) there should be no subtractions even if variable coefficients are *negative*, i.e. variables must be *forced positive*, (iv) interactions (variable products) should be supported. There are three possible types of two-variable interactions: categorical x continuous, categorical x categorical and continuous x continuous.

It is easier to use nomogram with variables that have maximum scores of 10 or 100. Variable rescaling can be achieved as follows. Let us suppose that the maximum score for any given variable is $M = 10$. Then, given a model with dummy variables with positive coefficients created for categorical variables, continuous variables with positive coefficients,

$$F = M / \max(x_i \cdot \alpha_i \quad i=1..N) \quad \forall \alpha_i$$

New variable coefficients can be then calculated as:

$$\varepsilon_i = \alpha_i \cdot F$$

The adjustment must be then also made in the *Total Score (TS)* term:

$$TS \cdot F = \left(\frac{p}{1-p} - \alpha_0 \right) \cdot F$$

Setting a boundary for probabilities can be done in the following manner:

$$p_{\min} = \frac{1}{1 + e^{-\alpha_0}}$$

$$p_{\max} = \frac{1}{1 + e^{-(\alpha_0 + \max(TS))}}$$

The *Total Score* is then adjusted to improve readability:

$$TS|_{p_{\max}} > TS|_{p=0.999} \Rightarrow TS'_{\max} = TS|_{p=0.999}$$

$$TS|_{p_{\max}} \leq TS|_{p=0.999} \Rightarrow TS'_{\max} = TS|_{p_{\max}}$$

Is it desirable to avoid subtractions on nomograms. Variables can be forced positive by generating a *different* regression with transformed variables. When a regression is executed in Stata, coefficients, variance-covariance matrices, gradient vectors and other parameters are stored in memory. Other commands can then make use of these parameters to make further calculations. Hence, executing a regression which was different from the one initially executed by the user could affect subsequent operations. This is why we implemented `nomolog` and `nomocox` as post-estimation commands and performed all modifications internally with variable transformations, without changing the regression command and its post-estimation outputs.

Given a regression with a single *categorical* variable A with N+1 categories and a regression constant α_0 ,

$$LP = \alpha_0 + TS = \alpha_0 + \alpha_{A1} \cdot A_1 + \alpha_{A2} \cdot A_2 + \dots + \alpha_{AN} \cdot A_N$$

where $A_1..A_N$ are dummies corresponding to each of the categories and categories are exclusive.

If $\exists \alpha_{Ai} \text{ } i=1..N < 0$, the most negative coefficient $\min(\alpha_{Ai} \text{ } i=1..N)$ is set as reference.

Then, since only one category is possible at a time,

$$LP = \alpha'_0 + \alpha'_{A1} \cdot A_1 + \alpha'_{A2} \cdot A_2 + \dots + \alpha'_{AN} \cdot A_N$$

where

$$\alpha'_0 = \alpha_0 + \min(\alpha_{Ai} \text{ } i=1..N)$$

$$\alpha'_{A1} = \alpha_{A1} - \min(\alpha_{Ai} \text{ } i=1..N)$$

...

$$\alpha'_{AN} = \alpha_{AN} - \min(\alpha_{Ai} \text{ } i=1..N)$$

Continuous variables can be transformed with a displacement of the variable range to positive values. These changes can be linearly combined for different variables, adjusting the constant term each time.

Although some researchers have stated that interactions cannot be introduced in nomograms [224], the opposite is true. As stated before, three kinds of two-variable interactions (products) are possible in **nomolog** and **nomocox**: categorical x categorical, categorical x continuous, continuous x continuous. The first kind results into categorical variables and the other two result in continuous variables. If an interaction between continuous variables is defined, one of them must be limited to a set of reference points, so that the nomogram remains linear.

Our nomogram generator supports both of the operators (# = interaction, ## = full interaction) used to produce interactions in modern version of Stata: x1#x2 and x1##x2 [228]. The first possibility (x1#x2) includes just the interaction term. The second one (x1##x2) includes the interaction term and the main effects (variables used in the interaction individually). Generally, when interactions are used, both the interaction and

the variables (main effects) are introduced in the model, due to the so-called “hierarchical principle”. We may explain this with the example that follows.

Let us suppose that we have two categorical variables: x_1 = Treatment, x_2 = Surgery; each of which can have two possible values:

- $x_1 = 0$ (Treatment A),
- $x_1 = 1$ (Treatment B)
- $x_2 = 0$ (Surgery not performed),
- $x_2 = 1$ (Surgery performed)

Let us then define a very simple model:

$$LP = \alpha_0 + \alpha_1 \cdot \text{Treatment} + \alpha_2 \cdot \text{Surgery} + \alpha_{12} \cdot \text{Treatment} \cdot \text{Surgery}$$

If only the main effects are included, i.e. $\alpha_1 \cdot \text{Treatment} + \alpha_2 \cdot \text{Surgery}$, all four combinations are different, but the difference in the outcome between patients subject to surgery is the same for both Treatment A and Treatment B. This may *not* be a reasonable assumption in some situations. Perhaps, the effect of surgery varies in combination with treatments A and B differently. If only the interaction term is included, i.e. $\text{Treatment} \cdot \text{Surgery} \cdot \alpha_{12}$, this implies that the difference between three of the possibilities ($x_1=1$ and $x_2=0$, $x_1=0$ and $x_2=1$, $x_1=0$ and $x_2=0$) are identical and only one of the possibilities ($x_1=1$ and $x_2=1$) is different, which may not be correct. The same derivation may be used within interactions of variables with several categories.

Given an interaction between two variables (X1 and X2), the scale of the first one is shown (X1) and various scales to the other (X2), one for each value of the first one. If both variables are categorical, X1 is the first variable in the command and X2 is shown for all values of X1. If both are continuous, the same criterion is used and the user must specify a maximum of 5

values of X1 for which X2 is represented. If one variable is continuous and the other is categorical, the categorical value is considered X1 and the other one X2 [205]. **nomolog** and **nomocox** allow displaying both the simplified and non-simplified versions of the nomogram.

A relevant problem is that if variable interactions are used, the variables participating in the interaction can no longer be forced positive with a simple transformation.

Continuing, with our example,

$$LP = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2$$

If we make the transformation $y_1 = A - x_1$, this will not only change coefficients α_1 and α_3 , but also coefficients α_0 and α_2 :

$$\begin{aligned} LP &= \beta_0 + \beta_1 y_1 + \beta_2 x_2 + \beta_3 y_1 x_2 = \beta_0 + \beta_1 y_1 + \beta_2 x_2 + \beta_3 y_1 x_2 = \beta_0 + \beta_1 (A - x_1) + \beta_2 x_2 + \beta_3 (A - x_1) x_2 = \\ &= \beta_0 + \beta_1 A - \beta_1 x_1 + \beta_2 x_2 + \beta_3 A x_2 - \beta_3 x_1 x_2 = \beta_0 + \beta_1 A - \beta_1 x_1 + (\beta_2 + \beta_3 A) x_2 - \beta_3 x_1 x_2 \end{aligned}$$

Therefore:

$$\alpha_0 = \beta_0 + \beta_1 A \Rightarrow \alpha_0 - \beta_1 A = \beta_0 = \alpha_0 + \alpha_1 A$$

$$\alpha_1 = -\beta_1 \text{ and } \alpha_3 = -\beta_3$$

$$\alpha_2 = \beta_2 + \beta_3 A \Rightarrow \alpha_2 - \beta_3 A = \beta_2 = \alpha_2 + \alpha_3 A$$

This is why, when interactions are used, the variables included in these interactions are *not* forced positive on the nomogram.

Examples of nomograms with interactions can be found in **Appendix C**.

4.1.4. Derivation of Cox regression nomograms

Once the logistic regression nomogram generator program (**nomolog**) was ready, we decided to develop another program for the generation of Cox regression nomograms (**nomocox**). There were two main reasons: (i) Cox regression models are very popular in modern clinical research; (ii) operations on the LP term defined for the generation of logistic regression nomograms could be adapted to Cox regression nomograms with ease.

The formula of a Cox regression can be written as:

$$S(t) = S_0(t) \cdot e^{(\alpha_1 x_1 + \dots + \alpha_N x_N)} = S_0(t) \cdot e^{LP} = S_0(t) \cdot e^{(TS)}$$

where

$S(t)$ = probability of survival for time $\geq t$

$S_0(t)$ = base survival

N = number of independent variables

$\alpha_{i=1..N}$ = independent variable coefficients

$x_{i=1..N}$ = independent variable values

LP = “linear predictor” = TS = *Total Score*

Conversely,

$$LP = \ln \left(\frac{S(t)}{S_0(t)} \right)$$

Notice that, in this case, there is no α_0 in the LP .

Variables with negative coefficients can be forced positive. Let us suppose that x_I is a variable with a negative coefficient. Then, the transformation $x_I = A - y$ results in

$$\begin{aligned} S(t) &= S_0(t) \cdot e^{(\alpha_1 x_1 + \dots + \alpha_N x_N)} = S_0(t) \cdot e^{(\alpha_1 (A-y) + \dots + \alpha_N x_N)} = S_0(t) \cdot e^{(\alpha_1 A - \alpha_1 y + \dots + \alpha_N x_N)} = \\ &= (S_0(t) \cdot e^{\alpha_1 A}) \cdot e^{-\alpha_1 y + \dots + \alpha_N x_N} \end{aligned}$$

In order to represent this function with linear nomograms, specific values (points of interest) of t need to be specified, i.e. Cox regression nomograms allow performing calculations for probabilities of survival for time $\geq t=T$.

$$S(t) \Big|_{t=T} = S_0(t) \cdot e^{(\alpha_1 x_1 + \dots + \alpha_N x_N)} \Big|_{t=T} = S_0(T) \cdot e^{(\alpha_1 x_1 + \dots + \alpha_N x_N)} = S_0(T) \cdot e^{(TS)}$$

where, $TS = Total\ Score =$ sum of individual variable scores.

It is also possible to display probability estimations for several time points of interest simultaneously. For instance, it may be possible to display *Total Score to Probability (of survival)* conversions for $T=12$ months, $T=24$ months and $T=60$ months simultaneously (please, see **Appendix C** for an example).

In the Cox regression, the base survival function $S_0(t)$ usually does not have a closed-form expression. Thus, if the nomogram has to be calculated for a T that does *not* have a value for $S_0(t=T)$, the closest value of t for which there is a value in $S_0(t)$ is chosen instead and a message is displayed notifying this. In other aspects, the operations described for the linear predictor (LP) are essentially the same as in the case of logistic regression nomograms.

4.1.5. Example of a nomogram for clinical practice: evolution of back pain

We may illustrate the use of our nomogram generator with an extract from a predictive model developed on a back pain registry of 17 centers of the Spanish National Health Service [230].

“Clinically relevant improvement of low back pain” was used as the dependent variable. Independent variables used in the model were “cronicity of back pain” (cronic, subacute, acute), “baseline severity of low back pain”, “baseline severity of referred pain”, “neuroreflexotherapy” (yes / no), “baseline grade of disability in Roland-Morris questionnaire score”, “previous surgery” (yes / no) [230]. A predictive model of improvement of low back pain is built using the **logit** command. A nomogram is then generated with the **nomolog** program (**Figure 31**) [205].

Figure 31. Nomogram for low back pain clinically relevant improvement.

4.2. User survey

4.2.1. Introduction

We decided to perform a survey of users of *nomolog* and *nomocox* in order to evaluate usability, user satisfaction, user perceptions of quality of documentation, and likelihood of recommendation to other potential users. We also included questions about the comparison between the exiting nomogram generators for the R statistical software and the ones we had developed in Stata. Other questions were included in order to assess the background of our users as researchers and users of nomograms.

Usability is defined as the “extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” in ISO 9241 part 11 [231]. However, there are no standards as to *how* exactly it can be evaluated. In specialized applications with educated users, it is usually reasonable to use customized rather than standardized questionnaires [232]. We were interested in a *summative* questionnaire, wherein users could evaluate their experience, rather than a questionnaire designed to detect usability problems. The latter could be achieved by means of direction communication with individual users. Given the highly specialized nature of the software, it was preferable to address usability problems and feature requests on an individual basis.

4.2.2. Materials and methods

4.2.2.1. Survey design

The questionnaire was intentionally short with only 21 questions (**Table 20**) in order to increase the number of respondents. We included several questions about the user's background: age, years as a researcher, highest educational achievement, industry or area of research. **nomolog** and **nomocox** were mostly used by healthcare researchers (biostatisticians, clinical practitioners, clinical researchers and epidemiologists). We did not know about the primary area of research of every respondent *a priori*, and included this variable in the questionnaire, in order to separate researchers in the areas of healthcare sciences from researchers from those in other fields.

We also inquired about knowledge of logistic and Cox regression procedures, since a basic understanding of both is necessary to use nomograms in a reasoned way.

We also included three questions referring to the specific usage a researcher could have given to logistic and Cox regression nomograms: (i) research (scientific publications, research reports, etc), (ii) clinical decision support (for instance, a graph showing a patient the probability of improvement after a certain clinical procedure), and (iii) teaching (for instance, showing students how to interpret logistic regression coefficients in a predictive model with a nomogram).

Since nomogram generation was possible with both Stata and the R statistical programming language, we thought relevant to ask about expertise with statistical software

in general, as well as the R statistical programming language and the Stata software in particular.

We also inquired about the number of times users had generated nomograms in Stata and in R. Even though nomogram generators in other software packages are rather limited, we also included a question about the number of executions of nomogram generators in other software packages, although we expected very few non-zero responses (with zero being equivalent to no executions).

We then asked about the ease of use and flexibility of nomogram generators in Stata and R, about the approximate time needed to generate the desired nomogram with **nomolog** or **nomocox** and about the overall satisfaction with the task. Finally, we included questions about the quality of documentation and the likelihood of recommendation to a friend or a colleague (using the Net Promoter Score [®], a metric which is widely used in business applications [233] and usability studies [232]).

<p>1. How old are you?</p> <ul style="list-style-type: none"> • 18-29 years • 30-49 years • 50-64 years • +65 years <p>2. What is your industry / area of research?</p> <p>Examples: Operations research, Medical statistics, Clinical doctor, etc.</p> <p>Answer:</p> <p>3. For how many years have you been in academic research?</p> <p>The answer must be a number. If you are not a researcher, answer 0.</p> <p>Answer (must be a number):</p> <p>4. What is your educational attainment (i.e. highest completed degree of education)?</p> <ul style="list-style-type: none"> • Ph.D. or equivalent degree (or Doctor of Science) • Master's or equivalent degree • Bachelor's or equivalent degree • Other (lower than the other options) <p>5. How would you rate your knowledge of the logistic regression?</p> <ul style="list-style-type: none"> • Very high • High • Average • Low • Very low <p>6. How would you rate your knowledge of the Cox regression?</p> <ul style="list-style-type: none"> • Very high • High • Average • Low • Very low <p>7. Have you ever built nomograms for research purposes? (with software tools or manually)</p> <ul style="list-style-type: none"> • Yes • No <p>8. Have you generated nomograms for decision support systems used in clinical practice? (with software tools or manually)</p> <p>[Answer only if you a researcher in the clinical research]</p> <p>Example: a nomogram for patients to decide whether to follow a treatment or not.</p> <ul style="list-style-type: none"> • Yes • No <p>9. Have you generated nomograms for teaching purposes? (with software tools or manually)</p> <p>Example: show students how to calculate the probability of an outcome with a logistic regression model.</p> <ul style="list-style-type: none"> • Yes • No <p>10. How would you rate your level of expertise with statistical software in general?</p> <ul style="list-style-type: none"> • Very high • High • Average • Low • Very low 	<p>11. How would you rate your level of expertise with Stata?</p> <ul style="list-style-type: none"> • Very high • High • Average • Low • Very low <p>12. How would you rate your level of expertise with R? (R is a statistical programming language)</p> <ul style="list-style-type: none"> • Very high • High • Average • Low • Very low <p>13. Approximately how many times have you executed a nomogram generator with R?</p> <p>If you have never used R for these purposes, answer 0.</p> <p>Answer (must be a number):</p> <p>14. Approximately how many times have you executed other software nomogram generators not based on Stata nor R?</p> <p>Answer (must be a number):</p> <p>15. Do you believe that nomolog and nomocox are easier to use than nomogram generators for R?</p> <ul style="list-style-type: none"> • Yes • No • I don't know / I am not sure <p>16. Do you believe that nomolog and nomocox are more flexible than nomogram generators for R?</p> <ul style="list-style-type: none"> • Yes • No • I don't know / I am not sure <p>17. Approximately how many times have you executed nomolog and nomocox?</p> <p>Answer (must be a number):</p> <p>18. How many minutes does it take you, on average, to generate the desired nomogram with nomolog / nomocox?</p> <p>Answer (must be a number):</p> <p>19. Overall, the task of generating a nomogram with nomolog / nomocox was:</p> <p>Answer (must be a number; 0=Very difficult; 10=Very easy):</p> <p>20. Overall, I am satisfied with the support information (messages, documentation, online documentation at http://www.zlotnik.net/stata/nomograms) for generating nomograms with nomolog / nomocox.</p> <p>Answer (must be a number; 0=Strongly disagree; 10=Strongly agree):</p> <p>21. How likely are you to recommend nomolog / nomocox to a friend or colleague? (Net Promoter Score ®)</p> <p>Answer (must be a number; 0=Not at all likely; 10=Extremely likely):</p>
--	---

Table 20. Questionnaire for users of the *nomolog* and *nomocox* programs.

4.2.2.2. Sample collection

Perhaps the most challenging task of this survey was locating a highly specialized group of users of **nomolog** and **nomocox**, who would also be willing to answer a short questionnaire. One option was to ask on the Statalist forum [234]. However, this would not have been the most efficient approach, since there are hundreds of messages every day and few persons read them all – a specific message could be easily “lost in the stream and repetitive messages used to bring attention to a certain topic are shunned and may be flagged as *spam*. Also, most people use Statalist to asking and responding questions about Stata itself or general statistical analyses. Even if some attention was achieved on StataList, potential respondents would not necessarily be motivated to actually use the software for clinical decision support or research purposes.

We thus decided to follow a different approach. We published an article in the Stata Journal about **nomolog**, which is a nomogram generator for the logistic regression [205]. The Stata Journal includes the software packages described in the articles, but is notoriously slow in publishing updates to these software packages, since these have to coincide with the next *printed issue* of the journal, and there are only four issues per year. We thus included a small annotation in the article stating that users could request the latest version contacting us by e-mail. Most importantly, Cox regression models are arguably more used in clinical research and we did not provide a download link for **nomocox**, a nomogram generator for the Cox regression (closely based on **nomolog**), since it was still a *beta* version. However, we added another annotation in the article stating that it could also be requested from the authors by e-mail.

We also did our best to optimize our website [235] for search engines and link all the presentations we did on the topic of nomograms [236, 237] with our site, in the hope that

potential users would contact us through the aforementioned website. Our assumption was that any user who would be willing to use the software in a serious way, both for research and clinical practice, would request the latest version of both packages. Therefore, in a note in the aforementioned article [205], we advised readers to visit our website to get the latest updates of both programs. This allowed us to compile a mailing list of people who had the potential to be active users of both software packages.

After a usage time of a minimum of one month after the submission of the programs (to give sufficient testing time), we contacted our users and asked them to complete a questionnaire. We implemented the survey on the Typeform platform [238], due to its user-friendliness and the possibility of generating unlimited surveys without payment. We created individual copies of the same questionnaire for each respondent, thus keeping track of the users who had actually answered the questionnaire. Those who did not respond were contacted weekly, with a maximum of four reminders. If the potential respondents did not answer after the fourth reminder, we stopped contacting them.

When we reached a minimum number of biomedical researchers ($N = 30$), we stopped seeking additional respondents.

4.2.2.3. Adjusted Wald interval for binary data

There are several ways of calculating confidence intervals for binary data. Perhaps the oldest method is the Laplace interval, generally known as Wald interval:

$$\hat{p} \pm z_{(1-\alpha/2)} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

where

$\hat{p}$ is the sample proportion

n is the sample size

$z_{(1-\alpha/2)}$ is the critical value from the normal distribution for the desired level of confidence

The problem with the Wald interval is that it's very inaccurate for small samples. More specifically, it is too *narrow*. In fact, with small samples, 95% Wald confidence intervals actually contain the real proportion far less than 95% of the cases [239]. Since this estimation is based on the Central Limit Theorem, a minimum sample size of 30 is required [232], but larger samples are preferable.

On the other hand, *exact* confidence intervals perform well for small sample sizes. These ensure that the real proportion is contained *at least* the percentage of cases specified by the confidence interval. Such is, for instance, the case of the Clopper-Pearson formula [240]. The main issue is that these estimations tend to produce excessively *wide* confidence intervals.

A less-known alternative is the *adjusted* Wald confidence interval, which provides a reasonable trade-off between the probability of inclusion of the actual proportion (*coverage*) and the width of confidence interval. It is computed as follows [232, 239]:

Step 1: compute $\hat{p}_{adj}$

$$\hat{p}_{adj} = \frac{x + z_{(1-\alpha/2)}^2 / 2}{n + z_{(1-\alpha/2)}^2}$$

where

x is number of “*successes*”

n is number of “*attempts*”

$z_{(1-\alpha/2)}$ is critical value from the normal distribution for the desired level of confidence

Step 2: use the original Wald formula to compute the confidence interval for p_{adj}

$$\hat{p}_{adj} \pm z_{(1-\alpha/2)} \sqrt{\frac{\hat{p}_{adj} \cdot (1 - \hat{p}_{adj})}{n}}$$

4.2.2.4. Geometric mean of time needed to complete a task (time-task data)

One of the best-known methods of measuring average values of continuous data with small samples is the t -distribution, which takes the size of the sample into account:

$$\bar{x} \pm t_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$$

where

$\bar{x}$ is the sample mean

s is the sample standard deviation

n is the sample size

$t_{(1-\alpha/2)}$ is the critical value from the t -distribution for the desired level of confidence, for $n-1$ degrees of freedom

However, when a distribution cannot be assumed to be symmetric around the mean, which often happens with “time taken to complete a task” (task-time) variables, other measurements should be considered. One possibility is the so-called *geometric mean* and its confidence interval, which has been shown to be a better estimate than the median for time-task measurements in small-sample usability studies [241]. It is calculated as follows:

Step 1: $\text{timeln} = \ln(\text{time})$

Step 2: $\text{avgtimeln} = \text{avg}(\text{timeln})$

Step 3: $\text{geometric mean} = e^{(\text{avgtimeln})}$

Confidence intervals can be obtained on the transformed value, i.e. $\ln(\text{time})$, and then converted back with exponentiation [232].

4.2.3. Results

A total of 37 questionnaires were sent. A total of 34 people answered the questionnaire (91,89% response rate). 3 users were discarded since they answered “0” to Q.17 (**Table 20**), and thus never used the **nomolog** and **nomocox** programs (these were two Psychologists and one clinical researcher). 1 user was discarded since he was not a biomedical researcher, and our primary objective was to analyze biomedical researchers. Had the sample size of non-clinical researchers been bigger, we could have analyzed these cases separately. Since it was so small, it was simply discarded. This left 30 questionnaires answered by clinical researchers, which would be used for general analyses (**Figure 32**). Among these, there were 16 who also used nomogram generators in R (Q.13).

Figure 32. Questionnaire response processing scheme.

A considerable geographic variety was observed among the respondents, with researchers from Australia, Asia, Europe and North America (**Table 21**).

Country	Institution	N
Australia	Institute for Health Research, University of Notre Dame	1
Australia	Royal Brisbane Hospital	1
Brazil	A. C. Camargo Cancer Center	1
China	Beijing Cancer Hospital	1
China	Zhejiang University School of Medicine Hospital	1
China	Modern Cancer Hospital Guangzhou	1
China	Nantong University	1
China	Peking University	1
China	Sun Yat-sen University Cancer Center	1
China	Chinese Academy of Medical Sciences & Peking Union Medical College	2
India	Tata Memorial Hospital	2
Italy	Fondazione IRCCS Istituto Nazionale Tumori	1
Italy	Università Vita-Salute San Raffaele	1
Korea	Seoul National University Hospital	1
Netherlands	University Medical Center Utrecht	1
Norway	Cancer Registry of Norway	1
Spain	Hospital Clínic & Universitat de Barcelona	1
Spain	Hospital Clínica Universitaria de Navarra	1
Spain	Hospital Universitario Puerta de Hierro-Majadahonda	1
Spain	Hospital Universitario Ramón y Cajal	3
Sweden	Karolinska Institute	1
Thailand	Mahidol Oxford Research Unit	1
UK	University of Glasgow	1
UK	University of Manchester	1
USA	Harvard Medical School & Massachusetts General Hospital	1
USA	University of South Florida	1

Table 21. Affiliations of the researchers who answered the questionnaire.

All respondents had at least B.Sc. degrees and most (60%) had Ph.D. or Doctor of Science degrees. 86,67% of the respondents had more than five years of experience as researchers. As for their areas of research, 46.67% were Biostatisticians, 13.33% were Epidemiologists and the remaining 40.00% were Clinical researchers (in the fields of Medical Oncology, Radiation Oncology and Nephrology) (**Table 22**). Among the Biostatisticians there were three *Chairs of Biostatistics*.

Almost all respondents had an average or higher expertise of both logistic and Cox regression models. All had an average or higher expertise with Stata (**Table 22**).

Age	N	%N
18-29 years	4	13,33%
30-49 years	19	63,33%
50-64 years	6	20,00%
+65 years	1	3,33%

Educational achievement	N	%N
Other (lower than the other options)	0	0,00%
Bachelor's or equivalent degree	2	6,67%
Master's or equivalent degree	10	33,33%
Ph.D. or equivalent degree (or Doctor of Science)	18	60,00%

Industry / Area of research	N	%N
Biostatistics	14	46,67%
Epidemiology	4	13,33%
Clinical researchers	12	40,00%

Years as a researcher	N	%N
1-2	4	13,33%
3-5	8	26,67%
6-10	7	23,33%
11-15	3	10,00%
16-20	4	13,33%
21-25	0	0,00%
26-30	2	6,67%
35+	2	6,67%

Knowledge of Cox regression	N	%N
Very low	0	0,00%
Low	1	3,33%
Average	8	26,67%
High	15	50,00%
Very high	6	20,00%

Knowledge of logistic regression	N	%N
Very low	0	0,00%
Low	1	3,33%
Average	8	26,67%
High	15	50,00%
Very high	6	20,00%

Knowledge of R	N	%N
Very low	8	26,67%
Low	14	46,67%
Average	5	16,67%
High	2	6,67%
Very high	1	3,33%

Knowledge of Stata	N	%N
Very low	0	0,00%
Low	0	0,00%
Average	11	36,67%
High	16	53,33%
Very high	3	10,00%

Table 22. Demographics, researcher profile, knowledge of logistic & Cox regression, knowledge of Stata, R and general statistical software.

Adjusted Wald 95% confidence intervals were used to compute estimations of proportions in the answers to questions about the usage they had given to logistic and Cox regression nomograms (**Table 23**): clinical decision support (44.43%-79.20%), research purposes (65.12%-93.98%) and teaching purposes (18.13%-52.32%). All but 3 respondents used nomograms for at least one of these purposes.

Have you ever used nomograms as clinical decision support?	N	%N
No	11	36,67%
Yes	19	63,33%

adj-Wald

95% LCI	%Yes	95% UCI
44,43%	61,82%	79,20%

Have you ever used nomograms for research purposes?	N	%N
No	5	16,67%
Yes	25	83,33%

adj-Wald

95% LCI	%Yes	95% UCI
65,12%	79,55%	93,98%

Have you ever used nomograms for teaching purposes?	N	%N
No	20	66,67%
Yes	10	33,33%

adj-Wald

95% LCI	%Yes	95% UCI
18,13%	35,22%	52,32%

Table 23. Usage given to logistic and Cox regression nomograms by the respondents. *LCI = Lower confidence interval; UCI = Upper confidence interval.*

The distribution of responses to Q.19, Q.20 and Q.21 can be seen in **Figure 33**. Confidence intervals based on the *t*-distribution were used for these continuous variables, with a range of 0 to 10 each. The average user experience (Q.19) was qualified as “Good-Excellent”, with a mean score of 8.73, 95% CI (8.33 - 9.14), on a scale from 0 to 10. The average quality of documentation was qualified as “Excellent”, with a mean score of 8.9, 95% CI (8.46 - 9.34) , on a scale from 0 to 10. The average user was very likely to recommend these programs to colleagues or friends, with a mean Net Promoter Score of 9.43, 95% CI (9.11 - 9.75).

The *geometric mean* of the time taken to generate a nomogram was (Q.18) was 5.66 minutes with a 95% CI (3.5, 9.1).

Q.19 - Overall, the task of generating a nomogram with nomolog / nomocox was
0 =Very difficult , 10 =Very easy

Q.20 – Overall, I am satisfied with the support information (messages, documentation, online documentation at <http://www.zlotnik.net/stata/nomograms>) for generating nomograms with nomolog / nomocox.
0 = Strongly disagree, 10 = Strongly agree

Q.21 - How likely are you to recommend nomolog / nomocox to a friend or colleague?
0 = Not at all likely, 10 = Extremely likely

Figure 33. Answers to Q.19, Q.20 and Q.21.

The comparison between Stata and R was carried out with a more limited dataset, $N=16$, as explained previously (**Figure 32**). Adjusted Wald confidence intervals were used.

100% of the respondents, with a 95% CI (75.83 %, 100%), believed that **nomocox** and **nomolog** were easier to use than the equivalent nomogram generators in R (Q.15). 81.25% of the respondents, with a 95% CI (54.03%, 96.36%) believed that **nomolog** and **nomocox** were more flexible than the equivalent programs in R (Q.16).

4.2.4. Discussion

Although sampling was opportunistic and the sample size was not large ($N=30$), we believe that it was representative, since it included the intended audience of researchers in the fields of Biostatistics, Epidemiology and Clinical Practitioners who used nomograms, with a wide geographic variety, advanced educational achievements, adequate understanding of logistic and Cox regression models, as well as statistical software. Most importantly, most of them used logistic and Cox nomograms for research purposes and many used them for clinical decision-making (**Table 22**). Among the respondents there were many experienced researchers affiliated with top international research institutions, such as Harvard Medical School, the Karolinska Institute or the Chinese Academy of Medical Sciences (**Table 21**).

We believe our results provide convincing evidence of a satisfying user experience, general ease of use and adequate documentation (**Figure 33**). Most respondents were “Promoters” (those with a Net Promoter Score > 8), i.e. those most likely to recommend the software to other researchers. The geometric mean needed to produce a nomogram was less than 10 minutes with a 95% CI, which is a very good result. This was most likely due to reasonable default options in both **nomolog** and **nomocox**.

The comparison of Stata nomogram generation with R was more limited ($N=16$) due to the exclusion of the respondents who had never used R for nomogram generation. All respondents (100%) found **nomocox** and **nomolog** easier to use, with a 95% CI (75.83 %, 100%). A raw proportion of 81.25%, with a 95% CI (54.03%, 96.36%), found our programs to be more flexible than the nomogram generators available in R.

4.3. Conclusions

We believe that we have achieved our objective of developing flexible and user-friendly programs to generate logistic and Cox regression nomograms [205].

Our implementations produce nomograms in the way they were conceived by Kattan and subsequently used in most publications – without confidence intervals. Displaying confidence intervals would make the nomograms cumbersome, since the standard error of the *Total Score* varies depending on individual variable values. We do not consider this to be a relevant limitation, since nomograms should be used for adequately calibrated and validated predictive models, for which confidence intervals are less important.

Interactions with three or more variables (example: `a##b##c`) are not supported, since these are very infrequent in biomedical research.

Although Stata possesses a rich graphical library, custom graphs are challenging to develop when hard-coded limitations are reached. In this case, time series graphs were used with the `xtline` command that imposes a limitation of 70 labels per graphic. In order to overcome this limitation, **nomolog** and **nomocox** implement a *pagination* procedure, which divides the nomogram in 10 variables per page. This can also be mitigated with the application of user-defined variable division numbers [205].

We believe that of our user survey results provide convincing evidence of: (i) the interest exhibited by experienced clinical researchers in these nomogram generators, given the numerous requests of the latest software versions; (ii) the correctness of the implementation, as it has undergone peer review in a high-impact journal [205] and has been subject to extensive testing by both the reviewers of Stata Journal [205], and by the

user community; (iii) the adequacy of the implementation in terms of ease-of-use, documentation and time required to complete the task of generating a nomogram, given the results of the questionnaire; (iv) better usability of **nomolog** and **nomocox** nomogram generators compared to those existing in R; (v) higher flexibility of **nomolog** and **nomocox** compared to existing software implementations in R.

5. Conclusions, Contributions and Future work

5.1. Conclusions

5.1.1. Emergency department census forecasting

The first topic addressed in this Ph.D. thesis was the usage of machine learning algorithms for the forecasting of ED census levels for the allocation of nursing personnel in the ED (based on *maximum* ED census forecasts) as a primary objective, and the allocation of resource that can be based on *average* ED census forecasts as a secondary objective. In both cases, our aim was to reduce major and medium underestimation situations with a cost increase of less than 10%, compared to usual resource allocation policies.

In the case of *maximum* ED census forecasts, M5P was the best choice for the reduction of major and medium nursing personnel understaffing risks, for all forecast horizons. Compared to the *usual staffing levels*, personnel planning with M5P could reduce major understaffing (>9 nurse) risks more than 10-fold (a reduction to ~1% with M5P compared to ~13% obtained with usual nurse staffing levels); and could reduce medium understaffing (7-9 nurses) risks approximately 3-fold (a reduction to ~3% with M5P compared to ~10% obtained with usual staffing levels). The usage of M5P also implied 5% - 6.1% increases in estimated nursing personnel costs (compared to the usual staffing levels), which are acceptable given the large understaffing risk reductions. This fulfilled our objective of significant reduction of major and medium understaffing risk with a less than 10% cost increase.

In the case of *average* ED census forecasts, once again, M5P was the best choice for the reduction of major and medium underestimation risks, for all forecast horizons. Relative risk reductions were similar to those of maximum ED census forecasts (more than 10-fold reduction in major underestimation and approximately 3-fold reduction in medium underestimation), compared to *usual resource allocation policies*. However, in this case, the usual resource allocation policies already had low risks of major and medium underestimation (~2% risk of major underestimation and ~3.7% risk of medium underestimation). On the other hand, most importantly, in this case (average ED census forecasts), M5P led to a *cost reduction* of more than 15% compared to usual resource allocation policies. This fulfilled our objective of significant reduction of major and medium underestimation risks (even though the absolute differences were not large, the differences between M5P and usual resource allocation policies were statistically significant, as 95% confidence intervals did not overlap, for all forecast horizons).

We believe that our results prove that it is possible to produce a decision support system for the allocation of ED nursing personnel based on time series models reliant on the M5P algorithm. Forecasts can be produced with sufficient advance. We have tested a variety of forecast horizons (1 week, 2 weeks, 4 weeks and 8 weeks).

5.1.2. Prediction of inpatient admission from the emergency department

Another objective of this Ph.D. thesis was the development of machine learning models for the prediction of probability of inpatient admission from the emergency department (ED) for patients of all age groups and all triage chief complaints (excepting “Pregnancy”). Our intention was to rely on a small number of variables, available in any ED setting equipped with the internationally validated and widely used Manchester Triage System (MTS). We

aimed to develop classifiers with adequate performance in terms of both discrimination and calibration (goodness-of-fit).

We used logistic regression (LR) models, artificial neural networks (ANN) models and models based on an ad hoc ensemble classifier that optimized calibration (it combined a base LR model with a tree of MTS chief complaints with LogitBoost on its leaves). A custom method was designed for the evaluation of the algorithms, with increasingly larger train sets and 12 consecutive test sets of approximately monthly length. Average areas under the ROC curve (AUROCs) and average H-L χ^2 were estimated with 95% confidence intervals (CIs) on the results obtained on the aforementioned test sets. Results were as follows. For LR models, average AUROC = 0.8531, 95% CI (0.8501, 0.8561); for ANN models, average AUROC = 0.8568, 95% CI (0.8531, 0.8606); and for ad hoc ensemble classifier models, average AUROC = 0.8635, 95% CI (0.8605, 0.8665). Confidence intervals of average AUROCs for LR and ad hoc ensemble classifier models did not overlap. Confidence intervals of average AUROCs for LR and ANN models slightly overlapped, although ANN models had higher AUROCs than LR models in all but one of the 12 test sets. Average H-L χ^2 were, respectively, 35.15, 95% CI (32.57, 37.73) for LR models; 10.47, 95% CI (7.78, 13.17) for ANN models and 11.4, 95% CI (9.10, 13.75) for ad hoc ensemble classifier models. Both ANN and ad hoc ensemble classifier models possessed better calibration than LR models, with H-L p-values > 0.05 in most experiments. All H-L χ^2 for all experiments had 8 degrees of freedom (i.e. the 10 groups of probability could be formed in all cases). Both ANNs and ad hoc ensemble classifiers fulfilled the optimization objectives of AUROC > 0.85 (i.e. lower 95% CI for AUROC > 0.85) and reduction of H-L χ^2 , with H-L p-values > 0.05 in 10 of the 12 experiments.

We believe that we have achieved our objective of building models suitable for an admission prediction system for all age groups and all chief complaints (excepting “Pregnancy”), based

on few variables available right after triage. It should be possible to obtain equivalent variables in the ED of any specialized care hospital which uses the Manchester Triage System. We believe that this approach for developing decision support systems for admission prediction can also be used with other triage systems, although some adaptations would be necessary. The models developed in this work system may be used for inpatient and ED bed management, fast-track inpatient admission, outbound transfer planning and rapid discharge from the ED, as well as an early warning system for clinical personnel in the ED and hospitalization wards.

The ad hoc ensemble classifier model has acceptable calibration and discrimination metrics on the test datasets, is fast to train even on large datasets, and does not require complex parameterization – the only parameter is the number of iterations in LogitBoost leaf models, that can simply be set to a reasonably high value (such as 200). Since LogitBoost has been shown to be resistant to overfitting, we believe that the overall model, with LogitBoost on its leaves, should be resistant to overfitting too. We believe that this approach can be considered reasonably safe and easy to use. The programming implementation of the ensemble classifier can be deployed in different software environments. There are two possibilities of deployment: a pure Java implementation based on Weka's application programming interface (API) or a Stata-based program with a web interface integrated with the Weka API.

Logistic regression models have slightly worse discrimination and calibration metrics than ad hoc ensemble classifier and ANN models, but can also be used, along with a nomogram, as a contingency procedure, in case of failure of predictive models reliant on information systems. The MTS may also be used without an information system, by means of following the decision flowcharts for each chief complaint.

5.1.3. Logistic and Cox regression nomograms

Another topic addressed in this Ph.D. thesis was the development of flexible and user-friendly software for the generation of nomograms for predictive models based on logistic or Cox regressions.

We believe that we have achieved this objective. We developed two programs: *nomolog* for the generation of logistic regression nomograms and *nomocox* for the generation of Cox regression nomograms. The results of a user survey provide convincing evidence of: (i) the interest shown by experienced clinical researchers in these nomogram generators, given the numerous requests of the latest versions of the programs we had developed; (ii) the correctness of the implementation, as it has undergone peer review in a high-impact journal and has been subject to extensive testing by a sizable and experienced user community; (iii) the adequacy of the implementation in terms of ease-of-use, documentation and time required to complete the task of generating a nomogram, given the results of the user survey; (iv) better usability of *nomolog* and *nomocox* nomogram generators compared to those existing in R; (v) higher flexibility of *nomolog* and *nomocox* compared to existing software implementations in R.

5.2. Contributions

5.2.1. Emergency department census forecasting

We developed time-series models for the forecasting of maximum emergency department (ED) census levels and average ED levels for the allocation of resources in the ED. Data were provided by the Ramón y Cajal University Hospital of Madrid, Spain.

We used an 8-hour granularity in our forecasts, since many resources in EDs are organized in 8-hour shifts. Nursing personnel should be adjusted to maximum ED census forecasts. Average ED census forecasts can be used for such needs as the allocation of non-clinical personnel, radiological or clinical laboratory resources in the ED. In both cases (average and maximum ED census forecasts), we wanted to reduce risks of major and medium underestimation of resources with a cost increase of less than 10%. Our primary objective was the forecasting of *maximum* ED census levels with these restrictions. Our secondary objective was the forecasting of *average* ED census levels with these same restrictions.

For both *maximum* and *average* ED census forecasts, the best results were obtained with time series based on the M5P algorithm.

In the case of *maximum* ED census forecasts, compared to the *usual staffing levels*, personnel planning with M5P could reduce major understaffing (>9 nurse) risks more than 10-fold (a reduction to ~1% with M5P compared to ~13% obtained with the usual nurse staffing levels); and could reduce medium understaffing (7-9 nurses) risks approximately 3-fold (a reduction to ~3% with M5P compared to ~10% obtained with the usual staffing levels). The usage of M5P also implied 5% - 6.1% increases in estimated nursing personnel

costs (compared to the usual staffing levels), which are acceptable given the large understaffing risk reductions.

In the case of average ED census forecasts, relative reductions of major and medium underestimation risks were similar to those of maximum ED census forecasts (more than 10-fold reduction in major underestimation and approximately 3-fold reduction in medium underestimation; both compared to the *usual resource allocation policies*). However, in this case, the usual resource allocation policies already had low risks of major and medium underestimation (~2% risk of major underestimation and ~3.7% risk of medium underestimation). On the other hand, most importantly, in this case (average ED census forecasts), M5P led to a cost reduction of >15% compared to the usual resource allocation policies.

The aforementioned results held for long forecast horizons (even for 8 week ahead forecasts), which enables the forecasting models reliant on M5P to be used as a decision support system for ED resource allocation, with an 8-hour granularity of forecasts taking into consideration underestimation risks, overestimation risks and costs of different resource allocation policies. To the best of our knowledge, this is the first such decision support system developed in Spain. Although some adjustment and more extensive testing are likely to be needed, we believe that our approach could also be used in other hospitals in Spain and in other countries.

5.2.2. Prediction of inpatient admission from the emergency department

We designed a system that can predict the probabilities of inpatient admission from the emergency department (ED), for patients of all age groups and all Manchester Triage System (MTS) chief complaints (excepting “Pregnancy”). The resulting predicted probabilities of inpatient admission can be used for such purposes as inpatient and ED bed management, fast-track inpatient admission, rapid discharge from the ED and outbound transfer planning, as well as an early warning system for clinical personnel in the ED and in the hospitalization wards. Data were provided by the Ramón y Cajal University Hospital of Madrid, Spain.

We designed logistic regression (LR) and artificial neural network (ANN) models, as well as models based on an ad hoc ensemble classifier. The latter consists of a LR model and a tree of LogitBoost submodels for MTS chief complaints.

We performed an evaluation with increasingly larger train sets and 12 consecutive test sets of approximately monthly length. Average areas under the ROC curve (AUROCs) and average H-L χ^2 were estimated with 95% confidence intervals (CIs) on the results obtained on the aforementioned test sets. All models (LR, ANN and ad hoc ensemble classifier) exhibited *adequate* discrimination and calibration (goodness-of-fit) for this classification problem, although ANN models outperformed LR models in terms of calibration and ad hoc ensemble classifier models outperformed LR models in terms of *both* discrimination and calibration.

Detailed results were as follows. For LR models, average AUROC = 0.8531, 95% CI (0.8501, 0.8561); for ANN models, average AUROC = 0.8568, 95% CI (0.8531, 0.8606) and for ad hoc

ensemble classifier models, average AUROC = 0.8635, 95% CI (0.8605, 0.8665). Confidence intervals of average AUROCs for LR and ad hoc ensemble classifier models did not overlap. Average H-L χ^2 were, respectively, 35.15, 95% CI (32.57, 37.73) for LR models; 10.47, 95% CI (7.78, 13.17) for ANN models and 11.4, 95% CI (9.10, 13.75) for ad hoc ensemble classifier models. Both ANN and ad hoc ensemble classifier models possessed better calibration than LR models, with H-L p-values > 0.05 in 10 of the 12 experiments.

ANN and ad hoc classifier models must be deployed on an information system with real-time integration with the hospital information systems. LR models with a nomogram representation can be used as a contingency procedure.

We believe that our models exhibit adequate performance given that they were designed for almost all patients visiting the ED of the Ramón y Cajal University Hospital. Our models rely exclusively on variables available in the EDs of most specialized care hospitals right after triage, which allows for real-time prediction of probabilities. To the best of our knowledge, this is the first such system designed in Spain. We believe that our approach can be adapted to other hospitals in Spain and in other countries.

5.2.3. Logistic and Cox regression nomograms

We developed two programs for the generation of logistic and Cox regression nomograms: *nomolog* and *nomocox*. These programs have been made available to the public and have been used by many experienced biomedical researchers, some of them from top academic institutions, such as the Karolinska Institute, the Harvard Medical School or the Chinese Academy of Medical Sciences. We surveyed our users about their background, their user experience with our nomogram generators, as well as the ease-of-use and flexibility of our programs compared to those available in the R statistical software. Most respondents were “Promoters” (Net Promoter Score > 8), i.e. very likely to recommend the software to other researchers. Respondents used nomograms for clinical decision support 95% CI (44.43%-79.20%), research purposes 95% CI (65.12%-93.98%) and teaching purposes 95% CI (18.13%-52.32%).

All respondents (100%) who had used both of our programs (*nomolog* and *nomocox*) and those available for the R statistical software, found *nomocox* and *nomolog* easier to use with a 95% adjusted Wald CI (75.83%, 100%). A raw proportion of 81.25%, with a 95% adjusted Wald CI (54.03%, 96.36%) found our programs to be more flexible than the nomogram generators available in the R statistical software.

Given these results, we believe that we have achieved our objective of developing software that enabled the generation of nomograms for logistic and Cox regression models in a flexible and user-friendly manner.

5.3. Future work

5.3.1. Emergency department census forecasting

We believe that performance could be further improved with time series models which could take into account ED census levels in *other* hospitals within the same region, since such models could be capable of capturing overall seasonality effects better. The inclusion of other variables, such as the volume of Internet visits to websites with clinical information [242] could also be considered in future models.

The performance of SARIMA models could be improved with the usage of several levels of seasonality. Procedures for automatically adjusting such SARIMA models will be explored in future research.

An unexplored consequence of a dynamic staffing policy would be the effect of staff handover to the next shift. Problems could arise with large staffing differences between subsequent shifts. Future research will address these situations.

ED throughput is dependent on many factors, such as available physicians, clinical laboratory throughput and radiological testing throughput. There could also be some effect in nurse staffing levels themselves. This could produce *feedback effects* between nurse staffing levels and ED census levels. Indeed, with higher nurse staffing levels, ED *throughput* could be increased and thus hour-by-hour ED census levels could *decrease*. Future research will address this hypothesis, most likely with simulation techniques.

The stratification of ED visits by severity and nursing workload should be addressed in future models, especially in settings where the majority of ED visits are *not* low severity.

However, we did not possess a normalized system for classifying patient visits based on their clinical severity and thus could not perform a severity-based analysis. Alternatively, analyses should be performed based on patient *location* in the ED, which could be used as a proxy of care intensity. In most ED settings in Spain hour by hour patient location *by ED area* (examples: consultations / blue / green area, yellow area, boxes, orange area, pre-admissions, etc.) is not available, but should be addressed in future ED information systems. Task-time measurements for nursing procedures could also be used for more sophisticated adjustments.

We have used an average estimation of personnel costs (gross costs of 125 euros per 8-hour shift per nurse). Actual personnel costs vary depending on many factors (type of contract, seniority, incentives, etc). More precise cost estimations will be considered in future research.

Our models can be interpreted as estimations of understaffing and overstaffing risks for nursing personnel. Hence, risk management approaches from other disciplines could be useful. For instance, perhaps a variation of the Hurst formula [243], used by Soviet engineers during the construction of the Aswan High Dam, could be used. Some ideas from modern financial risk management could also be adapted to ED staff planning [244]. Some adjustments are likely to be necessary if these approaches are followed since *strong shocks* (of 3 or more standard deviations) in other domains (such as some financial markets) are much more likely than in ED visits and thus ED census levels.

5.3.2. Prediction of inpatient admission from the emergency department

The usage of data from several hospitals could improve the ad hoc ensemble classifier in several ways: (i) more accurate submodels could be developed for each chief complaint; (ii) information about previous visits to *other* hospitals could be obtained; (iii) more extensive testing could be performed; (iv) outbound patient transfers would not have to be excluded from the analysis, since most would become inbound patient transfers to different hospitals in the same region; (v) it could be possible to account for the variability in admission patterns across different institutions; (vi) hospital clustering could be employed, then an additional sublevel in the “LogitBoost tree” model, corresponding to “hospital type”, could be introduced. Another possibility would be to use hospitals directly as another level of the tree. Obviously, the dataset must be reasonably uniform, for example, hospitals must use the same triage system or there should be a reasonable normalization procedure for chief complaints and triage levels.

Triage systems are not designed with admission prediction in mind. Additional clinical and sociological variables could be obtained as part of the triage process. Also, additional clinical variables could be collected for the most “error-prone” chief complaints and then included only in chief-complaint-specific submodels (**Appendix B**). For these, specific variables reliant on data of previous hospital visits could improve performance. As previously explained in **Section 3.3.4**, the usage of variables reliant on ICD coding and DRGs is problematic. We believe that these limitations could be at least partially overcome with Electronic Health Records that made extensive use of medical ontologies based on Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT). For instance, variables related to heart rate turbulence [245] detected in previous episodes of care could be used.

In settings with adequate integration of data between primary and specialized care information systems, variables describing chronic conditions (severe diabetes, chronic obstructive pulmonary disease, etc.) [25] could also improve predictive performance. In most developed countries, coding is done *during* patient encounters in primary care settings. The International Classification of Primary Care (ICPC) is a common choice for this kind of coding. In Spain, a local version called “Clasificación Internacional de la Atención Primaria, 2ª edición” (CIAP-2) is used. In the case of CIAP-2, information about clinical conditions can be obtained from components 70 to 99 (for example, chronic obstructive pulmonary disease is registered as chapter “R”, component “95”).

The inclusion of detailed information about previous visits to the emergency department along with information about chronic conditions (obtained from primary care) could result in a large number of correlated variables, which are likely to cause convergence problems in some algorithms. These problems can be attenuated with (i) a careful selection of only the most relevant chronic conditions; (ii) the usage of LogitBoost with weak learners; (iii) the usage of regularization techniques (such as penalized logistic regression).

Feature selection was performed based on expert knowledge and previous research. Triage chief complaint severity groups were derived based on the inspection of univariate effects on inpatient admission rates. Perhaps, performance could be further improved using raw datasets and automated feature extraction methodologies. However, we believe that much larger datasets are necessary for successful automated feature extraction, and, unlike in other problems such as image or speech recognition, where datasets are usually limited only by budgetary constraints, reasonably uniform datasets (same region, same hospital type, same triage type and same triage version) for inpatient admission prediction from the

emergency department cannot be expanded beyond the volume of clinical care actually provided to patients.

An unexplored path is the usage of data dimensionality reduction techniques, such as Singular value decomposition (SVD) or principal component analysis (PCA). However, as mentioned in **Section 3.3.3**, we did not see a reason to use these techniques for the set of independent variables we had chosen. Perhaps, if many correlated variables with no clear interpretation were considered in future models, dimensionality reduction techniques could be used. However, we believe that regularization or boosting (with weak learners) procedures would probably be better in terms of resistance to overfitting. One of the reasons is that dependent variable(s) should *not* be used in dimensionality reduction techniques such as PCA, thus, potentially useful information for the classification problem is lost when using these techniques. Furthermore, we believe that dimensionality reduction techniques could *worsen* calibration metrics, although this hypothesis should be subjected to experimental validation.

Alternative algorithms will be considered in future research. An initially promising SVM algorithm, which allowed for an explicit trade-off between *calibration* and *discrimination*, was the “Doubly-optimized Support Vector Machine” (DOC-SVM) [139]. Unfortunately, the implementation provided by the first author after contacting him, worked only with *perfectly balanced designs*, i.e. datasets where the number of positive samples is equal to the number of negative samples. Another relevant restriction was that this implementation was based on the CVX convex optimization numerical library [246], which, although very flexible, had a poor computational performance even for very small SVM problems. Finally, in the theoretical justification of the joint optimization framework, it was stated that the *minimization* of the Brier score [247] resulted in the improvement of calibration *and* the worsening of discrimination. A counterexample is a classifier with a

Brier score of 0, perfectly discriminative ($\text{AUROC} = 1$) and not calibrated at all (one such example can be found on page 649 of [248]). Further research is needed to assess the adequacy of DOC-SVM for prediction problems where both discrimination and calibration must be taken into account.

Future work will also address the usage of alternative ANN algorithms, such as Artificial Metaplasticity ANNs [160-163].

5.3.3. Logistic and Cox regression nomograms

The possibility of generating nomograms for Poisson regressions and *multinomial* logistic regressions will be considered in future versions.

6. Publications

6.1. International scientific journals (primary)

Computers Informatics Nursing (Q3 Medical Informatics / Q4 Nursing), 2016

Building a decision support system for inpatient admission prediction with the Manchester Triage System (MTS) and administrative check-in variables

Zlotnik A, Cuchi M, Perez-Perez MC, Gallardo-Antolín A, Montero-Martínez JM

URL:

http://journals.lww.com/cinjournal/Fulltext/2016/05000/Building_a_Decision_Support_System_for_Inpatient.6.aspx

Computers Informatics Nursing (Q3 Medical Informatics / Q4 Nursing), 2015

Emergency Department Visit Forecasting and Dynamic Nursing Staff Allocation Using Machine Learning Techniques With Readily Available Open-Source Software

Zlotnik A, Gallardo-Antolín A, Cuchi M, Perez-Perez MC, Montero-Martínez JM

URL:

http://journals.lww.com/cinjournal/Abstract/2015/08000/Emergency_Department_Visit_Forecasting_and_Dynamic.7.aspx

Stata Journal (Q1 Statistics), 2015

A general-purpose nomogram generator for predictive logistic regression models. Stata Journal. 2015. Volume 15, Number 2.

Zlotnik A, Abaira V

URL: <http://www.stata-journal.com/article.html?article=st0391>

6.2. International scientific journals (secondary)

Health Information Management Journal (Q3-Q4 Health Policy), 2015

Lifting the weight of a diagnosis-related groups family change: a comparison between refined and non-refined DRG systems for top-down cost accounting and efficiency indicators

Zlotnik A, Cuchi M, Perez-Perez MC

URL: <http://himaa2.org.au/HIMJ/sites/default/files/HIMJ1340Zlotnik.pdf>

6.3. International scientific conferences (primary)

IC-ININFO 2014

Calculating classifier calibration performance with a custom modification of Weka

Zlotnik A, Montero-Martínez JM, Gallardo-Antolín A

URL: <http://www-gth.die.upm.es/research/documentation/AG-150Cal-14.pdf>

2014 Spanish Stata Users Group Meeting: *two presentations*

1) Development of the nomolog program and its evolution: Toward the implementation of a nomogram generator for the Cox regression

Zlotnik A, Abraira V

URL: <http://www.stata.com/meeting/spain14/abstracts/>

2) A formal methodology for the comparison of results from different software packages: A case study of estimation of Hosmer–Lemeshow “deciles of risk” for a logistic regression with Stata and with a custom Java program

Zlotnik A, Montero-Martínez JM, Gallardo-Antolín A

URL: <http://www.stata.com/meeting/spain14/abstracts/>

2013 Spanish Stata Users Group Meeting

Stata logistic regression nomogram generator

Zlotnik A, Abraira V

URL: <http://www.stata.com/meeting/spain13/abstracts/>

HEALTHINF 2013

A Comparison of Multivariate SARIMA and SVM Models for Emergency Department Admission Prediction

Zlotnik A, Montero-Martínez JM, Gallardo-Antolín A

URL: http://www.healthinf.biostec.org/Abstracts/2013/HEALTHINF_2013_Abstracts.htm

6.4. International scientific conferences (secondary)

2015 Spanish Stata Users Group Meeting

Scientific organizer

2015 Spanish Stata Users Group Meeting

Stata web services: Toward Stata-based healthcare informatics applications integrated in a service-oriented architecture (SOA)

Zlotnik A, Escobar M, Gallardo-Antolín A, Montero-Martínez JM

URL: <http://www.stata.com/meeting/spain15/>

INTERSPEECH 2015

Random Forest-Based Prediction of Parkinson's Disease Progression Using Acoustic, ASR and Intelligibility Features

Zlotnik A, Montero-Martínez JM, San Segundo R, Gallardo-Antolín A

URL: <http://dblp2.uni-trier.de/rec/bibtex/conf/interspeech/ZlotnikMSG15>

URL2: [http://interspeech2015.org/wp-](http://interspeech2015.org/wp-content/uploads/direct/INTERSPEECH_2015_AbstractBook.pdf)

[content/uploads/direct/INTERSPEECH_2015_AbstractBook.pdf](http://interspeech2015.org/wp-content/uploads/direct/INTERSPEECH_2015_AbstractBook.pdf)

2015 UK Stata Users Group Meeting

Stata for Internet applications: A web interface for Stata user-written commands

Zlotnik A

URL: <http://www.stata.com/meeting/uk15/abstracts/>

References

- [1] "Ley 16/2003, de 28 de mayo, de cohesión y calidad del Sistema Nacional de Salud," ed: BOE (Boletín Oficial del Estado), 2003.
- [2] G. Lopez Casanovas, *et al.*, "Diversity and regional inequalities in the Spanish 'system of health care services'," *Health Economics*, vol. 14, pp. S221-S235, 2005.
- [3] Bloomberg.com. (2013. Available: <http://www.advisory.com/daily-briefing/2013/08/28/bloomberg-ranks-the-worlds-most-efficient-health-care-systems>
- [4] D. B. Evans, *et al.*, "Comparative efficiency of national health systems: cross national econometric analysis," *Bmj*, vol. 323, pp. 307-310, 2001.
- [5] A. Tandon, *et al.*, *Measuring overall health system performance for 191 countries*: World Health Organization, 2001.
- [6] A. Zlotnik, *et al.*, "Lifting the weight of a diagnosis-related groups family change: a comparison between refined and non-refined DRG systems for top-down cost accounting and efficiency indicators," *The HIM journal*, 2014.
- [7] *Ley 16/2012, de 27 de diciembre, por la que se adoptan diversas medidas tributarias dirigidas a la consolidación de las finanzas públicas y al impulso de la actividad económica* 2013.
- [8] "Recursos económicos del Sistema Nacional de Salud, 2003-2015," 2016-03-02 2016.
- [9] A. Castedo, *et al.*, "Decenas de municipios se quedan sin urgencias nocturnas en toda España," in *El País*, ed, 2013.
- [10] "Los achaques de la sanidad," in *Diario Información* ed, 2015.
- [11] "Sanidad desconoce cómo devolverá la Atención Primaria a los inmigrantes tres meses después de su anuncio " *eldiario.es*, 2015.
- [12] J. M. Pines, *et al.*, "International perspectives on emergency department crowding," *Academic Emergency Medicine*, vol. 18, pp. 1358-1370, 2011.
- [13] "Informe anual del Sistema Nacional de Salud," Ministerio de Sanidad, Servicios Sociales e Igualdad 2013.
- [14] A. Abellán García and R. Pujol Rodríguez, "El estado de la población mayor en España, 2015," 2015.
- [15] N. R. Hoot and D. Aronsky, "Systematic review of emergency department crowding: causes, effects, and solutions," *Annals of emergency medicine*, vol. 52, pp. 126-136, 2008.
- [16] M. Kennedy, *et al.*, "Review article: leaving the emergency department without being seen," *Emergency Medicine Australasia*, vol. 20, pp. 306-313, 2008.
- [17] D. B. Richardson, "The access-block effect: relationship between delay to reaching an inpatient bed and inpatient length of stay," *Medical Journal of Australia*, vol. 177, pp. 492-495, 2002.
- [18] J. C. Moskop, *et al.*, "Emergency department crowding, part 1—concept, causes, and moral consequences," *Annals of emergency medicine*, vol. 53, p. 605, 2009.
- [19] S. L. Bernstein, *et al.*, "The effect of emergency department crowding on clinically oriented outcomes," *Academic Emergency Medicine*, vol. 16, pp. 1-10, 2009.
- [20] I. Palanca-Sánchez, *et al.*, "Unidad de urgencias hospitalarias: Estándares y recomendaciones," *Informes, estudios e investigación. Madrid: Ministerio de Sanidad y Política Social*, 2010.
- [21] J. C. Brillman, *et al.*, "Triage: limitations in predicting need for emergent care and hospital admission," *Annals of emergency medicine*, vol. 27, pp. 493-500, 1996.
- [22] J. M. Kosowsky, *et al.*, "Can emergency department triage nurses predict patients' dispositions?," *The American journal of emergency medicine*, vol. 19, pp. 10-14, 2001.
- [23] S. D. Levine, *et al.*, "How well do paramedics predict admission to the hospital? A prospective study," *The Journal of emergency medicine*, vol. 31, pp. 1-5, 2006.
- [24] S. Lopez-Aguila, *et al.*, "Predictive model for emergency hospital admission and 6-Month readmission," *The American journal of managed care*, vol. 17, p. e348, 2011.

- [25] Y. Sun, *et al.*, "Predicting hospital admissions at emergency department triage using routine administrative data," *Acad Emerg Med*, vol. 18, pp. 844-50, Aug 2011.
- [26] A. Cameron, *et al.*, "A simple tool to predict admission at the time of triage," *Emergency Medicine Journal*, pp. emermed-2013-203200, 2014.
- [27] L. H. Aiken, *et al.*, "Implications of the California nurse staffing mandate for other states," *Health Services Research*, vol. 45, pp. 904-921, 2010.
- [28] R. L. Kane, *et al.*, "The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis," *Medical care*, vol. 45, pp. 1195-1204, 2007.
- [29] R. Paw, "Emergency department staffing in England and Wales, April 2007," *Emergency Medicine Journal*, vol. 25, pp. 420-423, 2008.
- [30] T. Serratt, "California's nurse-to-patient ratios, part 1: 8 years later, what do we know about nurse-level outcome?," *J Nurs Adm*, vol. 43, pp. 475-80, Sep 2013.
- [31] T. Serratt, "California's nurse-to-patient ratios, part 2: 8 years later, what do we know about hospital level outcomes?," *J Nurs Adm*, vol. 43, pp. 549-53, Oct 2013.
- [32] T. Serratt, "California's nurse-to-patient ratios, part 3: 8 years later, what do we know about nurse-level outcome?," *J Nurs Adm*, vol. 43, pp. 581-5, Nov 2013.
- [33] S. J. Weiss, *et al.*, "Comparison of the national emergency department overcrowding scale and the emergency department work index for quantifying emergency department crowding," *Academic Emergency Medicine*, vol. 13, pp. 513-518, 2006.
- [34] S. S. Jones, *et al.*, "An independent evaluation of four quantitative emergency department crowding scales," *Academic Emergency Medicine*, vol. 13, pp. 1204-1211, 2006.
- [35] U. Hwang, *et al.*, "Measures of crowding in the emergency department: a systematic review," *Academic Emergency Medicine*, vol. 18, pp. 527-538, 2011.
- [36] K. Raj, *et al.*, "National Emergency Department Overcrowding Study tool is not useful in an Australian emergency department," *Emergency Medicine Australasia*, vol. 18, pp. 282-288, 2006.
- [37] S. S. Jones, "Real-time demand forecasting in the emergency department," in *AMIA... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium*, 2007, p. 997.
- [38] S. S. Jones, *et al.*, "A multivariate time series approach to modeling and forecasting demand in the emergency department," *Journal of biomedical informatics*, vol. 42, pp. 123-139, 2009.
- [39] L. M. Schweigler, *et al.*, "Forecasting models of emergency department crowding," *Academic Emergency Medicine*, vol. 16, pp. 301-308, 2009.
- [40] G. Abraham, *et al.*, "Short-term forecasting of emergency inpatient flow," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, pp. 380-388, 2009.
- [41] N. R. Hoot, *et al.*, "Measuring and forecasting emergency department crowding in real time," *Annals of emergency medicine*, vol. 49, pp. 747-755, 2007.
- [42] M. Wargon, *et al.*, "A systematic review of models for forecasting the number of emergency department visits," *Emerg Med J*, vol. 26, pp. 395-9, Jun 2009.
- [43] Y. Sun, *et al.*, "Forecasting daily attendances at an emergency department to aid resource planning," *BMC Emergency Medicine*, vol. 9, p. 1, 2009.
- [44] R. Champion, *et al.*, "Forecasting emergency department presentations," *Australian Health Review*, vol. 31, pp. 83-90, 2007.
- [45] B. Y. Reis and K. D. Mandl, "Time series modeling for syndromic surveillance," *BMC Medical Informatics and Decision Making*, vol. 3, p. 2, 2003.
- [46] A. Diehl, *et al.*, "Use of calendar and weather data to predict walk-in attendance," *Southern medical journal*, vol. 74, p. 709, 1981.
- [47] R. Abdel-Aal and A. Mangoud, "Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis," *Computer Methods and Programs in Biomedicine*, vol. 56, pp. 235-247, 1998.

- [48] S. Au-Yeung, *et al.*, "Predicting patient arrivals to an accident and emergency department," *Emergency Medicine Journal*, vol. 26, pp. 241-244, 2009.
- [49] H. Batal, *et al.*, "Predicting patient visits to an urgent care clinic using calendar variables," *Academic Emergency Medicine*, vol. 8, pp. 48-53, 2001.
- [50] P. C. Milner, "Ten-year follow-up of ARIMA forecasts of attendances at accident and emergency departments in the Trent region," *Statistics in medicine*, vol. 16, pp. 2117-2125, 1997.
- [51] D. Tandberg and C. Qualls, "Time series forecasts of emergency department patient volume, length of stay, and acuity," *Annals of emergency medicine*, vol. 23, pp. 299-306, 1994.
- [52] I. Galan, *et al.*, "Short-term effects of air pollution on daily asthma emergency room admissions," *European Respiratory Journal*, vol. 22, pp. 802-808, 2003.
- [53] F. Ko, *et al.*, "Effects of air pollution on asthma hospitalization rates in different age groups in Hong Kong," *Clinical & Experimental Allergy*, vol. 37, pp. 1312-1319, 2007.
- [54] D. M. Stieb, *et al.*, "Air pollution and emergency department visits for cardiac and respiratory conditions: a multi-city time-series analysis," *Environ Health*, vol. 8, p. 25, 2009.
- [55] A. F. G. Ros, *et al.*, "Efecto mediático del fútbol en los servicios de traumatología," *Lecturas: Educación física y deportes*, p. 6, 2004.
- [56] R. S. L. Meca, *et al.*, "La programación televisiva no influye en la afluencia a urgencias," *enfermeria21*, 2010.
- [57] M. Gerolimetto, "ARIMA and SARIMA models," *Ca'Foscari University of Venice, Italy*, 2010.
- [58] G. E. Box, *et al.*, *Time series analysis: forecasting and control* vol. 734: John Wiley & Sons, 2011.
- [59] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*: OTexts, 2014.
- [60] D. C. Montgomery, *et al.*, *Introduction to time series analysis and forecasting*: John Wiley & Sons, 2015.
- [61] R. Sakia, "The Box-Cox transformation technique: a review," *The statistician*, pp. 169-178, 1992.
- [62] G. Bontempi, *et al.*, "Machine learning strategies for time series forecasting," in *Business Intelligence*, ed: Springer, 2013, pp. 62-77.
- [63] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International journal of forecasting*, vol. 22, pp. 443-473, 2006.
- [64] T. Bollerslev, *et al.*, "ARCH models," *Handbook of econometrics*, vol. 4, pp. 2959-3038, 1994.
- [65] L. Bauwens, *et al.*, "Multivariate GARCH models: a survey," *Journal of applied econometrics*, vol. 21, pp. 79-109, 2006.
- [66] R. J. Hyndman and Y. Khandakar, "Automatic time series for forecasting: the forecast package for R," Monash University, Department of Econometrics and Business Statistics 2007.
- [67] C. Blakely, "X-13ARIMA-SEATS."
- [68] R. B. Darlington. (1996, *A Regression Approach to Time Series Analysis*. Available: <http://node101.psych.cornell.edu/Darlington/series/series0.htm>
- [69] A. K. Palit and D. Popovic, *Computational intelligence in time series forecasting: theory and engineering applications*: Springer Science & Business Media, 2006.
- [70] N. K. Ahmed, *et al.*, "An empirical comparison of machine learning models for time series forecasting," *Econometric Reviews*, vol. 29, pp. 594-621, 2010.
- [71] G. P. Zhang, *et al.*, "A simulation study of artificial neural networks for nonlinear time-series forecasting," *Computers & Operations Research*, vol. 28, pp. 381-396, 2001.
- [72] E. Alpaydin, *Introduction to machine learning*: MIT press, 2014.

- [73] C.-J. Lu, *et al.*, "Financial time series forecasting using independent component analysis and support vector regression," *Decision Support Systems*, vol. 47, pp. 115-125, 2009.
- [74] N. Sapankevych and R. Sankar, "Time series prediction using support vector machines: a survey," *Computational Intelligence Magazine, IEEE*, vol. 4, pp. 24-38, 2009.
- [75] U. Thissen, *et al.*, "Using support vector machines for time series prediction," *Chemometrics and intelligent laboratory systems*, vol. 69, pp. 35-49, 2003.
- [76] X. Wu, *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, pp. 1-37, 2008.
- [77] L. Breiman, *et al.*, "CART: Classification and regression trees," *Wadsworth: Belmont, CA*, vol. 156, 1983.
- [78] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 14-23, 2011.
- [79] J. Quinlan, "Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Australia (pp. 343-348)," ed: Singapore: World Scientific, 1992.
- [80] A. Polumetla, "Machine learning methods for the detection of RWIS sensor malfunctions," University of Minnesota, 2006.
- [81] A. Bakhach, *et al.*, "Forecasting Directional Changes in Financial Markets," 2015.
- [82] R. Urraca, *et al.*, "Estimation of solar global irradiation in remote areas," *Journal of Renewable and Sustainable Energy*, vol. 7, p. 023136, 2015.
- [83] S. Wang, *et al.*, "Application of nonparametric regression in predicting traffic incident duration," *Transport*, pp. 1-10, 2015.
- [84] E. Frank, *et al.*, "Alternating model trees," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, 2015, pp. 871-878.
- [85] Q. Duan, *et al.*, "Analysis and Prediction of Enterprise Service-Level Performance," in *Data-Driven Optimization and Knowledge Discovery for an Enterprise Information System*, ed: Springer, 2015, pp. 115-138.
- [86] E. Frank, *et al.* (2015, 2016-02-05). *Documentation for the AlternatingModelTree class*. Available: <http://weka.sourceforge.net/doc.packages/alternatingModelTrees/>
- [87] B. E. Boser, *et al.*, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144-152.
- [88] P. Rivas-Perea, *et al.*, "Support vector machines for regression: a succinct review of large-scale and linear programming formulations," 2012.
- [89] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*: Cambridge university press, 2000.
- [90] B. Schölkopf and A. Smola, "Support Vector Machines," *Encyclopedia of Biostatistics*, 1998.
- [91] A. Smola and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155-161, 1997.
- [92] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, pp. 199-222, 2004.
- [93] E. Alpaydin, "Introduction to machine learning," 2004.
- [94] R. Stefan, "SVM kernels for time series analysis," in *Proc. of Tagungsband der GI Workshop-Woche*, 2001, pp. 43-50.
- [95] (2016, 2016-02-07). *Weka API documentation: Class PolyKernel*. Available: <http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/supportVector/PolyKernel.html>
- [96] C. Soguero-Ruiz, *et al.*, "On the differential benchmarking of promotional efficiency with machine learning modelling (II): Practical applications," *Expert Systems with Applications*, vol. 39, pp. 12784-12798, 2012.

- [97] A. Al-Anazi and I. Gates, "Support vector regression for porosity prediction in a heterogeneous reservoir: A comparative study," *Computers & Geosciences*, vol. 36, pp. 1494-1503, 2010.
- [98] W. Zhang, *et al.*, "Forecasting of turbine heat rate with online least squares support vector machine based on gravitational search algorithm," *Knowledge-Based Systems*, vol. 39, pp. 34-44, 2013.
- [99] S. Shamshirband, *et al.*, "Wind turbine power coefficient estimation by soft computing methodologies: comparative study," *Energy Conversion and Management*, vol. 81, pp. 520-526, 2014.
- [100] S. Gupta and M. Starr, *Production and Operations Management Systems*: CRC Press, 2014.
- [101] A. Zlotnik, *et al.*, "Emergency Department Visit Forecasting and Dynamic Nursing Staff Allocation Using Machine Learning Techniques With Readily Available Open-Source Software," *Computers Informatics Nursing*, vol. 33, pp. 368-377, 2015.
- [102] M. Floyd, *QlikView Scripting*: Packt Publishing Ltd, 2013.
- [103] C. M. Christensen, *et al.*, "Will disruptive innovations cure health care?," *Harvard Business Review*, vol. 78, pp. 102-112, 2000.
- [104] R. J. Holden, "Lean Thinking in emergency departments: a critical review," *Annals of emergency medicine*, vol. 57, pp. 265-278, 2011.
- [105] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, pp. 1124-1131, 1974.
- [106] D. Kahneman, *Thinking, fast and slow*: Macmillan, 2011.
- [107] E. Molins, *et al.*, "Association between radiologists' experience and accuracy in interpreting screening mammograms," *BMC Health Services Research*, vol. 8, p. 91, 2008.
- [108] D. L. Miglioretti, *et al.*, "When Radiologists Perform Best: The Learning Curve in Screening Mammogram Interpretation 1," *Radiology*, vol. 253, pp. 632-640, 2009.
- [109] C. C. Gard, *et al.*, "Misclassification of Breast Imaging Reporting and Data System (BI-RADS) Mammographic Density and Implications for Breast Density Reporting Legislation," *The breast journal*, vol. 21, pp. 481-489, 2015.
- [110] H. G. Welch and H. J. Passow, "Quantifying the benefits and harms of screening mammography," *JAMA internal medicine*, vol. 174, pp. 448-454, 2014.
- [111] H.-D. Cheng, *et al.*, "Computer-aided detection and classification of microcalcifications in mammograms: a survey," *Pattern recognition*, vol. 36, pp. 2967-2991, 2003.
- [112] K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging," *The British Journal of Radiology*, 2014.
- [113] S. Wang and R. M. Summers, "Machine learning and radiology," *Medical image analysis*, vol. 16, pp. 933-951, 2012.
- [114] K. Levi, "Expert systems should be more accurate than human experts: evaluation procedures from human judgement and decision making," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 19, pp. 647-657, 1989.
- [115] N. Handly, *et al.*, "Evaluation of a hospital admission prediction model adding coded chief complaint data using neural network methodology," *European Journal of Emergency Medicine*, vol. 22, pp. 87-91, 2015.
- [116] E. Tsui, *et al.*, "Development of an automated model to predict the risk of elderly emergency medical admissions within a month following an index hospital visit: A Hong Kong Experience," *Health informatics journal*, vol. 21, pp. 46-56, 2015.
- [117] J. Wang, *et al.*, "Patient admission prediction using a pruned fuzzy min-max neural network with rule extraction," *Neural Computing and Applications*, vol. 26, pp. 277-289, 2015.
- [118] P. N. Skandalakis, *et al.*, "'To afford the wounded speedy assistance': Dominique Jean Larrey and Napoleon," *World journal of surgery*, vol. 30, pp. 1392-1399, 2006.

- [119] I. Robertson-Steel, "Evolution of triage systems," *Emergency Medicine Journal*, vol. 23, pp. 154-155, 2006.
- [120] R. C. Bone, *et al.*, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine," *Chest Journal*, vol. 101, pp. 1644-1655, 1992.
- [121] K. Mackway-Jones, *et al.*, *Emergency Triage. Manchester Triage Group*, 2nd ed.: Wiley Blackwell, 2006.
- [122] R. S. Bermejo, *et al.*, "El triaje en urgencias en los hospitales españoles," *Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias*, vol. 25, pp. 66-70, 2013.
- [123] G. E. d. T. Manchester. (2015, 2016-01-02). *El Sistema de Triage Manchester*. Available: http://www.triagemanchester.com/web/presentacion_es_66.php
- [124] T. R. M. Azeredo, *et al.*, "Efficacy of the Manchester Triage System: a systematic review," *International emergency nursing*, 2014.
- [125] I. Gräff, *et al.*, "The German Version of the Manchester Triage System and Its Quality Criteria—First Assessment of Validity and Reliability," *PloS one*, vol. 9, p. e88995, 2014.
- [126] I. Van der Wulp, *et al.*, "Predicting admission and mortality with the Emergency Severity Index and the Manchester Triage System: a retrospective observational study," *Emergency Medicine Journal*, vol. 26, pp. 506-509, 2009.
- [127] F. M. Sánchez, *et al.*, "Modelo predictor de ingreso hospitalario a la llegada al servicio de urgencias," in *Anales del Sistema Sanitario de Navarra*, 2012, pp. 207-217.
- [128] C. M. E. Martínez, "Modelo predictor del ingreso hospitalario desde el servicio de urgencias," PhD, Departamento de Medicina Preventiva Universidad Rey Juan Carlos, Madrid, 2012.
- [129] J. Crane and C. Noon, *The definitive guide to emergency department operational improvement: employing lean principles with current ED best practices to create the "no wait" department*: CRC Press, 2011.
- [130] D. G. Altman, *et al.*, "Prognosis and prognostic research: validating a prognostic model," *Bmj*, vol. 338, 2009.
- [131] R. Entezari-Maleki, *et al.*, "Comparison of classification methods based on the type of attributes and sample size," *Journal of Convergence Information Technology*, vol. 4, pp. 94-102, 2009.
- [132] S. Huang, *et al.*, "Quantitative prediction of acute ischemic tissue fate using support vector machine," *Brain research*, vol. 1405, pp. 77-84, 2011.
- [133] S. Y. Peng, *et al.*, "Random forest can predict 30 day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination," *European Journal of Neurology*, vol. 17, pp. 945-950, 2010.
- [134] S. H. S. A. Ubaidillah, *et al.*, "Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study," *Jurnal Teknologi*, vol. 65, 2013.
- [135] L. Xu, *et al.*, "A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery," *Remote Sensing of Environment*, vol. 141, pp. 14-23, 2014.
- [136] Q. Yuan, *et al.*, "[SVM-aided cancer diagnosis based on the concentration of the macroelement and microelement in human blood]," *Sheng wu yi xue gong cheng xue za zhi= Journal of Biomedical Engineering= Shengwu Yixue Gongchengxue Zazhi*, vol. 24, pp. 513-518, 2007.
- [137] I. Cohen and M. Goldszmidt, "Properties and benefits of calibrated classifiers," in *Knowledge Discovery in Databases: PKDD 2004*, ed: Springer, 2004, pp. 125-136.
- [138] A. Zlotnik, *et al.*, "Calculating classifier calibration performance with a custom modification of Weka," presented at the 4th International Conference on Integrated Information (IC-INNFO), Madrid, Spain, 2014.

- [139] X. Jiang, *et al.*, "Doubly Optimized Calibrated Support Vector Machine (DOC-SVM): An algorithm for joint optimization of discrimination and calibration," 2012.
- [140] D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression*: John Wiley & Sons, 2004.
- [141] J. Bartlett. (2014, 2014-06-09). *The Hosmer-Lemeshow goodness of fit test for logistic regression*. Available: <http://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/>
- [142] T. Cook. (2012, *Stat 642, Lecture notes for 04/12/05*. Available: <https://www.biostat.wisc.edu/~cook/642.tex/notes0412.pdf>
- [143] D. W. Hosmer and S. Lemeshow, "Goodness of fit tests for the multiple logistic regression model," *Communications in statistics-Theory and Methods*, vol. 9, pp. 1043-1069, 1980.
- [144] G. Bertolini, *et al.*, "One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model," *Journal of epidemiology and biostatistics*, vol. 5, pp. 251-253, 2000.
- [145] L. Á. O. Ochoa, *Gestión clínica: desarrollo e instrumentos*: Ediciones Díaz de Santos, 2012.
- [146] M. Vuk and T. Curk, "ROC curve, lift chart and calibration plot," *Metodoloski zvezki*, vol. 3, pp. 89-108, 2006.
- [147] P. D. Allison, "Measures of fit for logistic regression," in *Proceedings of the SAS Global Forum 2014 Conference*, 2014.
- [148] S. Boyd and L. Vandenberghe, *Convex optimization*: Cambridge university press, 2004.
- [149] J. Friedman, *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, pp. 337-407, 2000.
- [150] (2016, 2016-02-07). *Weka API documentation: Class LogitBoost*. Available: <http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/LogitBoost.html>
- [151] R. A. McDonald, *et al.*, "An empirical comparison of three boosting algorithms on real data sets with artificial class noise," in *Multiple Classifier Systems*, ed: Springer, 2003, pp. 35-44.
- [152] M. Schonlau, "Boosted regression (boosting): An introductory tutorial and a Stata plugin," *Stata Journal*, vol. 5, p. 330, 2005.
- [153] M. Taneja, *et al.*, "Prediction of click frauds in mobile advertising," in *Contemporary Computing (IC3), 2015 Eighth International Conference on*, 2015, pp. 162-166.
- [154] J. Ramírez, *et al.*, "Ensemble Tree Learning Techniques for Magnetic Resonance Image Analysis," in *Innovation in Medicine and Healthcare 2015*, ed: Springer, 2016, pp. 395-404.
- [155] J. Lorenzo Trueba, *et al.*, "Development of a genre-dependent TTS system with cross-speaker speaking-style transplantation," 2014.
- [156] B. Yegnanarayana, *Artificial neural networks*: PHI Learning Pvt. Ltd., 2009.
- [157] A. K. Jain, *et al.*, "Artificial neural networks: A tutorial," *Computer*, pp. 31-44, 1996.
- [158] J. Schürmann, *Pattern classification: a unified view of statistical and neural approaches*: Wiley Online Library, 1996.
- [159] A. Marciano-Cedeno and D. Andina, "Data mining for the diagnosis of type 2 diabetes," in *World Automation Congress (WAC), 2012*, 2012, pp. 1-6.
- [160] A. Marciano-Cedeno, *et al.*, "A prediction model to diabetes using artificial metaplasticity," in *New Challenges on Bioinspired Applications*, ed: Springer, 2011, pp. 418-425.
- [161] A. Marciano-Cedeño, *et al.*, "WBCD breast cancer database classification applying artificial metaplasticity neural network," *Expert Systems with Applications*, vol. 38, pp. 9573-9579, 2011.
- [162] A. Marciano-Cedeño, *et al.*, "Breast cancer classification applying artificial metaplasticity algorithm," *Neurocomputing*, vol. 74, pp. 1243-1250, 2011.

- [163] A. Marcano-Cedeno, *et al.*, "Artificial metaplasticity neural network applied to credit scoring," *International journal of neural systems*, vol. 21, pp. 311-317, 2011.
- [164] J. Macías-Guarasa, *et al.*, "Acoustical and lexical based confidence measures for a very large vocabulary telephone speech hypothesis-verification system," in *Proc. of ICSLP*, 2000.
- [165] N. Karayiannis and A. N. Venetsanopoulos, *Artificial neural networks: learning algorithms, performance evaluation, and applications* vol. 209: Springer Science & Business Media, 2013.
- [166] M. A. Nielsen, "Neural Networks and Deep Learning," *URL: <http://neuralnetworksanddeeplearning.com/>*.(visited: 01.11. 2014), 2015.
- [167] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, pp. 115-133, 1943.
- [168] R. Edward, "An Introduction to Neural Networks," *A White paper. Visual Numerics Inc., United States of America*, 2004.
- [169] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological review*, vol. 65, p. 386, 1958.
- [170] B. L. Kalman and S. C. Kwasny, "Why tanh: choosing a sigmoidal function," in *Neural Networks, 1992. IJCNN., International Joint Conference on*, 1992, pp. 578-581.
- [171] H. R. Maier and G. C. Dandy, "The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study," *Environmental Modelling & Software*, vol. 13, pp. 193-209, 1998.
- [172] C. M. Bishop, *Neural networks for pattern recognition*: Oxford university press, 1995.
- [173] J. J. Hopfield, "Learning algorithms and probability distributions in feed-forward and feed-back networks," *Proceedings of the National Academy of Sciences*, vol. 84, pp. 8429-8433, 1987.
- [174] K. Zhu, *et al.*, "Predicting 30-day Hospital Readmission with Publicly Available Administrative Database," *Methods of information in medicine*, vol. 54, pp. 560-567, 2015.
- [175] A. Zlotnik, *et al.*, "Building a decision support system for inpatient admission prediction with the Manchester Triage System (MTS) and administrative check-in variables," *Computers Informatics Nursing*, vol. 34, pp. 224-230, 2016.
- [176] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.
- [177] K. A. Hunt, *et al.*, "Characteristics of frequent users of emergency departments," *Annals of emergency medicine*, vol. 48, pp. 1-8, 2006.
- [178] V. Jairath, *et al.*, "Mortality From Acute Upper Gastrointestinal Bleeding in the United Kingdom: Does It Display a "Weekend Effect" &quest," *The American journal of gastroenterology*, vol. 106, pp. 1621-1628, 2011.
- [179] A. L. Sharp, *et al.*, "Don't get sick on the weekend: an evaluation of the weekend effect on mortality for patients visiting US EDs," *The American journal of emergency medicine*, vol. 31, pp. 835-837, 2013.
- [180] F. Maggs and M. Mallet, "Mortality in out-of-hours emergency medical admissions--more than just a weekend effect," *The journal of the Royal College of Physicians of Edinburgh*, vol. 40, pp. 115-118, 2010.
- [181] R. Busse, *et al.*, *Diagnosis-Related Groups in Europe: Moving towards transparency, efficiency and quality in hospitals*: McGraw-Hill Education (UK), 2011.
- [182] 3M, *All-Patient Refined Diagnosis Related Groups (APR-DRGs) Methodology Overview, version 24*, 2007.
- [183] A. Zlotnik, *et al.*, "Comparación de cinco años de casuística de hospitalización y ambulatoria con IR-GRD y AP-GRD en un hospital universitario español de alta especialización de 1.100 camas.," *Portales Médicos*, vol. X, 2015.
- [184] S. Peiró, "Métodos de evaluación de casuística y ajuste de severidad de riesgos," in *Gestión clínica: desarrollo e instrumentos*, ed: Ediciones Díaz de Santos, 2012.

- [185] (2015, 2016-03-05). *De AP-GRD a APR-GRD: implicaciones en la complejidad de los hospitales*. Available: <http://www.aimfa.es/de-ap-grd-a-apr-grd-implicaciones-en-la-complejidad-de-los-hospitales/>
- [186] B. Li, *et al.*, "Risk adjustment performance of Charlson and Elixhauser comorbidities in ICD-9 and ICD-10 administrative databases," *BMC Health Services Research*, vol. 8, p. 12, 2008.
- [187] D. A. Southern, *et al.*, "Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data," *Medical care*, vol. 42, p. 355, 2004.
- [188] I. Goldstein, *et al.*, "Three approaches to automatic assignment of ICD-9-CM codes to radiology reports," in *AMIA Annual Symposium Proceedings*, 2007, p. 279.
- [189] L. Kevers and J. Medori, "Symbolic classification methods for patient discharge summaries encoding into ICD," in *Advances in Natural Language Processing*, ed: Springer, 2010, pp. 197-208.
- [190] S. Nitsuwat and W. Paoín, "Development of ICD-10-TM ontology for a semi-automated morbidity coding system in Thailand," *Methods of information in medicine*, vol. 51, p. 519, 2012.
- [191] W. Paoín and S. Nitsuwat, "Development of Experience Base Ontology to Increase Competency of Semi-automated ICD-10-TM Coding System," *Information Technology Journal*, vol. 8, pp. 95-99, 2012.
- [192] S. Pereira, *et al.*, "Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding," in *MIE*, 2006, pp. 845-850.
- [193] M. Subotin and A. R. Davis, "A System for Predicting ICD-10-PCS Codes from Electronic Health Records," *ACL 2014*, p. 59, 2014.
- [194] N. Sánchez-Maróño, *et al.*, "Filter methods for feature selection—a comparative study," in *Intelligent Data Engineering and Automated Learning-IDEAL 2007*, ed: Springer, 2007, pp. 178-187.
- [195] D. G. Altman and P. K. Andersen, "Bootstrap investigation of the stability of a Cox regression model," *Statistics in medicine*, vol. 8, pp. 771-783, 1989.
- [196] S. Derksen and H. Keselman, "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables," *British Journal of Mathematical and Statistical Psychology*, vol. 45, pp. 265-282, 1992.
- [197] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267-288, 1996.
- [198] J. P. Leigh, "Assessing the importance of an independent variable in multiple regression: is stepwise unwise?," *Journal of clinical epidemiology*, vol. 41, pp. 669-677, 1988.
- [199] J. Doménech and J. Navarro, C. a. f. Stata., Ed., ed: Universitat Autònoma de Barcelona, 2015.
- [200] H. Demuth and M. Beale, "Neural network toolbox for use with MATLAB," 1993.
- [201] Mathworks. (2015, 2016-01-05). *scatteredInterpolant class*, *MATLAB Documentation*. Available: <http://www.mathworks.com/help/matlab/ref/scatteredinterpolant-class.html>
- [202] ---. (2015, 2016-01-07). *SimpleCart*. Available: <http://wiki.pentaho.com/display/DATAMINING/SimpleCart>
- [203] W. Gould, *et al.*, *Maximum likelihood estimation with Stata*: Stata press, 2006.
- [204] E. R. DeLong, *et al.*, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837-845, 1988.
- [205] A. Zlotnik and V. Abaira, "A general-purpose nomogram generator for predictive logistic regression models," *Stata Journal*, vol. Volume 15, Number 2, 2015.
- [206] M. A. LaMantia, *et al.*, "Predicting hospital admission and returns to the emergency department for elderly patients," *Academic Emergency Medicine*, vol. 17, pp. 252-259, 2010.

- [207] A. Zlotnik, "Stata for Internet applications: A web interface for Stata user-written commands," presented at the United Kindom Stata Users Group Meeting, London, United Kindom, 2015.
- [208] A. Zlotnik, *et al.*, "Stata web services: Toward Stata-based healthcare informatics applications integrated in a service-oriented architecture (SOA)," presented at the Spanish Stata Users Group Meeting, Madrid, Spain, 2015.
- [209] M. Escobar, "Studying coincidences with network analysis and other multivariate tools," *Stata Journal*, vol. 15, pp. 1118-1156, 2015.
- [210] J. López-Picazo Ferrer, *et al.*, "Estratificación y monitorización de niveles de adecuación de ingresos urgentes en un hospital de tercer nivel," *Revista de Calidad Asistencial*, vol. 29, pp. 10-16, 2014.
- [211] S. S. Chen, *et al.*, "Factors that influence the accuracy of triage nurses' judgement in emergency departments," *Emergency Medicine Journal*, vol. 27, pp. 451-455, 2010.
- [212] N. Parenti, *et al.*, "A systematic review on the validity and reliability of an emergency department triage scale, the Manchester Triage System," *International journal of nursing studies*, 2014.
- [213] H. Evesham, "Origins and development of nomography," *Annals of the History of Computing*, vol. 8, pp. 324-333, 1986.
- [214] G. Khovanskii, "Nomography and Its Possibilities," 1972.
- [215] G.S.Jovanovsky, *Nomography and its applications*: Naúka, 1977.
- [216] D. A. Grimes, "The nomogram epidemic: resurgence of a medical relic," *Annals of internal medicine*, vol. 149, pp. 273-275, 2008.
- [217] D. Lin and T. R. Fleming, *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis: Survival Analysis* vol. 123: Springer Science & Business Media, 2012.
- [218] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 187-220, 1972.
- [219] M. W. Kattan, *et al.*, "A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer," *Journal of the National Cancer Institute*, vol. 90, pp. 766-771, 1998.
- [220] M. W. Kattan and J. Marasco, "What is a real nomogram?," in *Seminars in oncology*, 2010, pp. 23-26.
- [221] M. W. Kattan, *et al.*, "A postoperative prognostic nomogram for renal cell carcinoma," *The Journal of urology*, vol. 166, p. 63, 2001.
- [222] M. W. Kattan and P. T. Scardino, "Prediction of progression: nomograms of clinical utility," *Clinical prostate cancer*, vol. 1, pp. 90-96, 2002.
- [223] M. W. Kattan, *et al.*, "Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer," *Journal of Clinical Oncology*, vol. 17, pp. 1499-1499, 1999.
- [224] F. K. H. Chun, *et al.*, "A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer," *BJU international*, vol. 99, pp. 794-800, 2007.
- [225] F. E. Harrell, *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*: Springer, 2001.
- [226] F. Harrell, "rms: Regression modeling strategies. R package version 3.4-0," ed, 2012.
- [227] D. Yang, "Build Prognostic Nomograms for Risk Assessment Using SAS® " in *SAS Global Forum 2013*, 2013.
- [228] L. StataCorp StataCorp, "College Station, TX: 2011," *Stata statistical software: release*, vol. 12.
- [229] C. F. Baum, "An introduction to Stata programming," *Stata Press books*, 2009.
- [230] F. M. Kovacs, *et al.*, "Predicting the evolution of low back pain patients in routine clinical practice: results from a registry within the Spanish National Health Service," *Spine J*, vol. 12, pp. 1008-20, Nov 2012.

- [231] ISO, "Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) Part 11: Guidance on Usability (ISO 9241-11:1998(E))," ed. Geneva: ISO, 1998.
- [232] J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*: Elsevier, 2012.
- [233] F. F. Reichheld, *et al.*, *The ultimate question: Driving good profits and true growth*: Harvard Business School Press Boston, MA, 2006.
- [234] StataCorp. (2016, *Statalist* | *The Stata Forum*. Available: <http://www.statalist.org/>
- [235] A. Zlotnik. (2015, 2015-12-15). *Nomograms in Stata*. Available: <http://www.zlotnik.net/stata/nomograms>
- [236] A. Zlotnik and V. Abairra, "Stata logistic regression nomogram generator," presented at the Spanish Stata Users Group Meeting, Madrid, Spain, 2013.
- [237] A. Zlotnik and V. Abairra, "Development of the nomolog program and its evolution: Toward the implementation of a nomogram generator for the Cox regression," presented at the Spanish Stata Users Group Meeting, Barcelona, Spain, 2014.
- [238] Typeform. (2016, 2015-12-20). *Typeform*. Available: <https://www.typeform.com/>
- [239] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, pp. 119-126, 1998.
- [240] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, pp. 404-413, 1934.
- [241] J. Sauro and J. Lewis, "Average task times in usability tests," in *The 28th international conference*, 2010, p. 2347.
- [242] A. Ekström, *et al.*, "Forecasting emergency department visits using internet data," *Annals of emergency medicine*, vol. 65, pp. 436-442. e1, 2015.
- [243] H. E. Hurst, "Long-term storage capacity of reservoirs," *Trans. Amer. Soc. Civil Eng.*, vol. 116, pp. 770-808, 1951.
- [244] E. Monte, *et al.*, "On the self-similar distribution of the emergency ward arrivals time series," *Fractals-an Interdisciplinary Journal on the Complex Geometry*, vol. 10, pp. 413-428, 2002.
- [245] C. Soguero-Ruiz, *et al.*, "Ontology for heart rate turbulence domain from the conceptual model of SNOMED-CT," *Biomedical Engineering, IEEE Transactions on*, vol. 60, pp. 1825-1833, 2013.
- [246] M. Grant, *et al.*, "CVX: Matlab software for disciplined convex programming," ed, 2008.
- [247] G. Blattenberger and F. Lad, "Separating the Brier score into calibration and refinement components: A graphical exposition," *The American Statistician*, vol. 39, pp. 26-32, 1985.
- [248] M. Woodward, *Epidemiology: Study Design and Data Analysis*: CRC Press, 2013.

Appendix A. Average ED census forecasts compared to personnel needs indicated by maximum ED census forecasts

In this Appendix, we provide an analysis of errors if *average* ED census forecasts were used for predicting actually needed nursing personnel (adjusted to *maximum* ED census levels).

As a summary, in this scenario, understaffing situations are reduced most with the *usual staffing levels*. However, *usual staffing levels* also imply a much higher incidence of overstaffing situations and higher estimated yearly costs than other algorithms.

Algorithm	Forecast horizon	MAE	% Major (>9 nurses) understaffing 95% CI	% Medium (7-9 nurses) understaffing 95% CI	% Minor (4-6 nurses) understaffing 95% CI	% Minor (4-6 nurses) understaffing 95% CI	% Medium (7-9 nurses) understaffing 95% CI	% Major (>9 nurses) understaffing 95% CI	Estimated yearly monetary cost (€)
Usual staffing	1 week	22.73	(12.38%, 14.39%)	(9.38%, 11.18%)	(13.31%, 15.38%)	(11.93%, 13.92%)	(6.52%, 8.06%)	(3.65%, 4.85%)	3,863,812.50
SA	1 week	18.69	(21.82%, 24.32%)	(16.10%, 18.34%)	(19.76%, 22.17%)	(3.76%, 4.97%)	(0.97%, 1.64%)	(0.14%, 0.46%)	3,357,156.25
SARIMA (S=1 week)	1 week	17.04	(20.28%, 22.71%)	(19.34%, 21.73%)	(23.14%, 25.69%)	(1.43%, 2.22%)	(0.29%, 0.71%)	(0.03%, 0.25%)	3,308,187.50
SARIMA (S=4 weeks)	1 week	17.98	(22.81%, 25.34%)	(18.98%, 21.36%)	(20.17%, 22.59%)	(1.90%, 2.80%)	(0.37%, 0.82%)	(0.07%, 0.34%)	3,289,500.00
Linear reg.	1 week	12.35	(15.50%, 17.71%)	(24.81%, 27.41%)	(32.31%, 35.11%)	(0.10%, 0.40%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,255,468.75
SVR	1 week	12.30	(17.40%, 19.70%)	(26.07%, 28.71%)	(30.75%, 33.52%)	(0.10%, 0.40%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,220,562.50
AMT	1 week	11.45	(18.09%, 20.42%)	(21.89%, 24.39%)	(26.63%, 29.29%)	(0.29%, 0.71%)	(0.00%, 0.15%)	(0.00%, 0.00%)	3,269,250.00
M5P	1 week	11.38	(13.99%, 16.11%)	(25.30%, 27.92%)	(32.04%, 34.84%)	(0.15%, 0.49%)	(0.00%, 0.00%)	(0.00%, 0.15%)	3,283,156.25
Usual staffing	2 weeks	22.73	(12.38%, 14.39%)	(9.38%, 11.18%)	(13.31%, 15.38%)	(11.93%, 13.92%)	(6.52%, 8.06%)	(3.65%, 4.85%)	3,863,812.50
SA	2 weeks	18.72	(21.80%, 24.29%)	(16.17%, 18.41%)	(19.63%, 22.03%)	(3.80%, 5.02%)	(0.97%, 1.64%)	(0.14%, 0.46%)	3,357,593.75
SARIMA (S=1 week)	2 weeks	18.17	(21.44%, 23.92%)	(18.94%, 21.31%)	(21.28%, 23.76%)	(1.74%, 2.60%)	(0.44%, 0.93%)	(0.14%, 0.46%)	3,303,343.75
SARIMA (S=4 weeks)	2 weeks	20.09	(25.71%, 28.34%)	(16.79%, 19.07%)	(18.33%, 20.68%)	(3.61%, 4.80%)	(0.93%, 1.59%)	(0.28%, 0.68%)	3,290,750.00
Linear reg.	2 weeks	12.50	(15.68%, 17.90%)	(25.12%, 27.73%)	(31.66%, 34.45%)	(0.17%, 0.52%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,254,687.50
SVR	2 weeks	12.37	(17.62%, 19.93%)	(26.38%, 29.03%)	(30.48%, 33.24%)	(0.10%, 0.40%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,217,125.00
AMT	2 weeks	11.58	(15.81%, 18.04%)	(24.29%, 26.87%)	(30.75%, 33.52%)	(0.22%, 0.60%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,267,250.00
M5P	2 weeks	11.55	(14.19%, 16.32%)	(24.65%, 27.25%)	(32.04%, 34.84%)	(0.17%, 0.52%)	(0.00%, 0.00%)	(0.00%, 0.15%)	3,285,406.25
Usual staffing	4 weeks	22.73	(12.38%, 14.39%)	(9.38%, 11.18%)	(13.31%, 15.38%)	(11.93%, 13.92%)	(6.52%, 8.06%)	(3.65%, 4.85%)	3,863,812.50
SA	4 weeks	18.78	(21.91%, 24.41%)	(16.06%, 18.29%)	(19.56%, 21.96%)	(3.78%, 4.99%)	(0.99%, 1.66%)	(0.14%, 0.46%)	3,356,156.25
SARIMA (S=1 week)	4 weeks	20.10	(24.72%, 27.32%)	(17.55%, 19.86%)	(18.67%, 21.03%)	(2.30%, 3.27%)	(0.69%, 1.27%)	(0.12%, 0.43%)	3,273,156.25
SARIMA (S=4 weeks)	4 weeks	22.61	(30.39%, 33.15%)	(14.57%, 16.72%)	(15.93%, 18.15%)	(4.61%, 5.94%)	(1.80%, 2.68%)	(0.55%, 1.09%)	3,252,875.00
Linear reg.	4 weeks	12.61	(16.08%, 18.32%)	(25.03%, 27.64%)	(31.66%, 34.45%)	(0.12%, 0.43%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,248,093.75
SVR	4 weeks	12.47	(17.91%, 20.24%)	(26.45%, 29.10%)	(30.26%, 33.01%)	(0.09%, 0.37%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,212,562.50
AMT	4 weeks	11.80	(16.15%, 18.39%)	(24.83%, 27.43%)	(31.88%, 34.68%)	(0.15%, 0.49%)	(0.00%, 0.15%)	(0.00%, 0.00%)	3,250,875.00
M5P	4 weeks	11.71	(14.44%, 16.58%)	(24.63%, 27.22%)	(31.84%, 34.63%)	(0.14%, 0.46%)	(0.00%, 0.07%)	(0.00%, 0.11%)	3,283,531.25
Usual staffing	8 weeks	22.73	(12.38%, 14.39%)	(9.38%, 11.18%)	(13.31%, 15.38%)	(11.93%, 13.92%)	(6.52%, 8.06%)	(3.65%, 4.85%)	3,863,812.50
SA	8 weeks	18.90	(21.96%, 24.46%)	(16.13%, 18.36%)	(19.49%, 21.89%)	(3.76%, 4.97%)	(0.99%, 1.66%)	(0.14%, 0.46%)	3,353,531.25
SARIMA (S=1 week)	8 weeks	22.31	(30.28%, 33.04%)	(15.81%, 18.04%)	(16.64%, 18.90%)	(3.23%, 4.36%)	(1.11%, 1.82%)	(0.26%, 0.66%)	3,182,562.50
SARIMA (S=4 weeks)	8 weeks	24.46	(32.56%, 35.37%)	(13.24%, 15.31%)	(14.13%, 16.25%)	(5.20%, 6.59%)	(2.32%, 3.30%)	(0.73%, 1.33%)	3,233,968.75
Linear reg.	8 weeks	12.86	(16.97%, 19.25%)	(25.37%, 27.99%)	(30.42%, 33.17%)	(0.07%, 0.34%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,235,718.75
SVR	8 weeks	12.63	(18.65%, 21.01%)	(26.34%, 28.99%)	(29.87%, 32.62%)	(0.03%, 0.25%)	(0.00%, 0.11%)	(0.00%, 0.00%)	3,203,687.50
AMT	8 weeks	11.79	(19.09%, 21.47%)	(23.98%, 26.55%)	(29.74%, 32.48%)	(0.28%, 0.68%)	(0.03%, 0.25%)	(0.00%, 0.07%)	3,223,531.25
M5P	8 weeks	11.90	(15.06%, 17.24%)	(24.56%, 27.15%)	(31.61%, 34.40%)	(0.14%, 0.46%)	(0.00%, 0.07%)	(0.00%, 0.11%)	3,275,968.75

Table A1. Results on the macro test set (4 years) with algorithms for *average* ED census forecasts used to predict *maximum* ED census levels. *95% CI overlapping with those of M5P are highlighted in bold.*

Figure A1. Estimated risks of understaffing(average ED census forecasts used to predict *maximum* ED census) for different algorithms and forecast horizons.

Figure A2. Estimated risks of overstaffing (*average* ED census forecasts used to predict *maximum* ED census) for different algorithms and forecast horizons.

Figure A3. Estimated yearly nursing personnel costs (*average ED census forecasts*) for different algorithms and forecast horizons.

Appendix B. Error distributions for inpatient admission prediction for the ad hoc ensemble classifier

Chief complaint group "S0"

Chief complaint group "S1"

Chief complaint group "S2"

Chief complaint groups "S3" and "S4"

Appendix C. Nomogram usage

How is a logistic regression nomogram used?

Step 1) Establish scores for all variable values.

Step 2) Obtain the *Total Score* by means of adding up all the scores obtained in the previous step.

Step 3) Obtain the probability of event (Total Score $\Rightarrow$ Probability of event).

Example:

Given an adequately calibrated logistic regression model and its nomogram (figure below), what is the probability of union affiliation for a worker who has a tenure of 7.5 years, works 30 hours per week, is 50 years old, is a college graduate and white?

Step 1) Establish scores for all variable values.

- *Job tenure (years) = 7.50 $\Rightarrow$ Score ≈ 3*
- *Usual hours worked = 30 $\Rightarrow$ Score ≈ 2*
- *Age in current year = 50 $\Rightarrow$ Score ≈ 2.5*
- *College grad $\Rightarrow$ Score ≈ 4.2*
- *Race = white $\Rightarrow$ Score ≈ 0*

Step 2) Obtain the Total Score adding up all the scores obtained in the previous step.

$$\text{Total score} = 3 + 2 + 2.5 + 4.2 + 0 = 11.7$$

Step 3) Obtain the probability of event (Total Score => Probability of event).

Total score = 11.7 is equivalent to a probability of approximately 0.28-0.29

How is a Cox regression nomogram used?

Step 1) Establish scores for all variable values.

Step 2) Obtain the Total Score by means of adding up all the scores obtained in the previous step.

Step 3) Obtain the probabilities of time-to-event (Total Score => Probability of time-to-event).

Example:

Given an adequately calibrated Cox regression model and its nomogram (figure below), what is the probability of survival for $t \geq 12$ months, for $t \geq 24$ months and for $t \geq 60$ months for an individual with no Ganglia, high LDH, V1=400 and V2=25?

Step 1) Establish scores for all variable values.

- *Ganglia = no* => *Score* ≈ 0

- *LDH = high* => *Score* ≈ 1.5

- *V1 = 900* => *Score* ≈ 9

- *V2 = 125* => *Score* ≈ 4.75

Step 2) Obtain the Total Score adding up all the scores obtained in the previous step.

$$\text{Total score} = 0 + 1.5 + 9 + 4.75 = 15.25$$

Step 3) Obtain the probability of event (Total Score => Probability of event).

Total score = 15.25 => probability of survival for $t \geq 12$ months of ≈ 0.8

Total score = 15.25 => probability of survival for $t \geq 24$ months of 0.6-0.7

Total score = 15.25 => probability of survival for $t \geq 60$ months of 0.3-0.4

Example of logistic regression nomogram with categorical x categorical interaction

Example:

Given an adequately calibrated logistic regression model and its nomogram (figure above), what is the probability of union affiliation for a worker who is 50 years old, is a college graduate and white?

Step 1) Establish scores for all variable values.

- Age in current year = 60 $\Rightarrow$ Score ≈ 8.5
- Race = white & College grad $\Rightarrow$ Score ≈ 5.5
- Race = white $\Rightarrow$ Score ≈ 0

Step 2) Obtain the Total Score by means of adding up all the scores obtained in the previous step.

$$\text{Total Score} = 8.5 + 5.5 + 0 = 14$$

Step 3) Obtain probability.

$$\text{Total Score} = 14 \Rightarrow \text{Probability of event between } 0.3 \text{ and } 0.4$$

Example of logistic regression nomogram with categorical x continuous interaction

Example:

Given an adequately calibrated logistic regression model and its nomogram (figure above), what is the probability of union affiliation for a worker who is 60 years old, has an annual salary of 20 (thousand dollars) and is of other race (neither white nor black)?

Step 1) Establish scores for all variable values.

- Age in current year = 60 $\Rightarrow$ Score ≈ 1
- Race = other & Wage = 20 $\Rightarrow$ Score ≈ -2.1
(remember that negative scores are possible in interactions)
- Race = other $\Rightarrow$ Score ≈ 2.8

Step 2) Obtain the Total Score by means of adding up all the scores obtained in the previous step.

$$\text{Total Score} = 1 - 2.1 + 2.8 = 1.7$$

Step 3) Obtain probability.

$$\text{Total Score} = 1.7 \Rightarrow \text{Probability of event} \approx 0.2$$

Example of logistic regression nomogram with continuous x continuous interaction

Given an adequately calibrated regression model and its nomogram (figure above), what is the probability of union affiliation for a worker who is 40 years old and has an annual salary of 30 (thousand dollars), and works 40 hours (a week)?

Step 1) Establish scores for all variable values.

- *Hours = 40 & Wage = 30 $\Rightarrow$ Score ≈ 3.4*
- *Age in current year = 40 $\Rightarrow$ Score ≈ 0.5*
- *Hours = 40 $\Rightarrow$ Score ≈ 2.1*

Step 2) Obtain the Total Score by means of adding up all the scores obtained in the previous step.

$$\text{Total Score} = 3.4 + 0.5 + 2.1 = 6$$

Step 3) Obtain probability.

$$\text{Total Score} = 6 \Rightarrow \text{Probability of event} \approx 0.64$$

Appendix D. Software developed for this Ph.D. thesis

- **nomolog & nomocox: nomogram generators for Stata.**
- **A web interface for Stata (similar to RServe).**
- **MATLAB – Weka integration layer.**
- **Stata – Weka integration layer.**
- **Modification of the –allsets– Stata command to allow the computation of evaluation metrics on validation or train datasets.**
- **Modified version of Weka for the computation of Hosmer-Lemeshow calibration metrics.**
- **A package for the calculation of Hosmer-Lemeshow calibration metrics for MATLAB.**
- **A data management system for the extraction, combination, validation and transformation of datasets commonly used in Emergency Departments.**