Class Imbalance Impact on the Prediction of Complications during Home Hospitalization: A Comparative Study

Mireia Calvo, Isaac Cano, Carme Hernández, Vicent Ribas, Felip Miralles, Josep Roca and Raimon Jané, *Senior Member*, *IEEE*

***Abstract*— Home hospitalization (HH) is presented as a healthcare alternative capable of providing high standards of care when patients no longer need hospital facilities. Although HH seems to lower healthcare costs by shortening hospital stays and improving patient's quality of life, the lack of continuous observation at home may lead to complications in some patients. Since blood tests have been proven to provide relevant prognosis information in many diseases, this paper analyzes the impact of different sampling methods on the prediction of HH outcomes. After a first exploratory analysis, some variables extracted from routine blood tests performed at the moment of HH admission, such as hemoglobin, lymphocytes or creatinine, were found to unmask statistically significant differences between patients undergoing successful and unsucessful HH stays. Then, predictive models were built with these data, in order to identify unsuccessful cases eventually needing hospital facilities. However, since these hospital admissions during HH programs are rare, their identification through conventional machine-learning approaches is challenging. Thus, several sampling strategies designed to face class imbalance were herein overviewed and compared. Among the analyzed approaches, over-sampling strategies, such as ROSE (Random Over-Sampling Examples) and conventional random over-sampling, showed the best performances. Nevertheless, further improvements should be proposed in the future so as to better identify those patients not benefiting from HH.**

I. INTRODUCTION

Home hospitalization (HH) emerged in response to the growing demand for hospital care and the high costs associated with the diagnosis and treatment of acute and chronic decompensated diseases. This healthcare alternative is capable of providing high standards of care through a set of home-based medical and nursing services, when patients no longer need hospital facilities, but still require active and complex surveillance. Indeed, HH has demonstrated to lower healthcare-associated costs by shortening hospital stays and avoiding readmissions; and it has been presented as an opportunity to improve integrated care [1], [2], [3]. Moreover, it reduces the risk of nosocomial infections, and it has been reported to improve patient's quality of life, by extending stays in a familiar environment [5], [6].

However, some relevant disease-related early signs may not be detected due to the lack of continuous observation at home, leading to complications that may result in regular hospital admissions. Thus, eligibility criteria in order to identify those patients who may not benefit from HH is of major importance. Since blood test data has proven to be a significant indicator of prognosis in many diseases [7], [8], [9], we hypothesized that those patients not being eligible for HH programs could be identified through routine blood tests, performed at the moment of admission.

Nevertheless, unsuccessful HH cases are rare. Since most conventional machine-learning approaches are extremely sensitive to class imbalance, they show a strong bias toward the majority class (far superior in number). To tackle this limitation, a number of sampling strategies have been developed, but their suitability is generally problem-dependent. Thus, after an exploratory statistical analysis of the available data to evaluate its potential to distinguish between successful and unsuccessful HH stays, this paper provides an overview of different sampling strategies and compares their impact on classification performance, in this particular scenario.

II. MATERIALS AND METHODS

A. Study population

The blood test data from 1951 patients having been admitted to the HH program delivered by the Integrated Care Unit at Hospital Clínic de Barcelona (Spain) was collected in the context of a prospective study conducted from January 2012 to December 2015, including the following variables:

- **L1300**: Leukocyte count (10^9/L), $[4.00 - 11.00]$
- **L1301**: Platelet count (10^9/L), $[130 - 400]$
- **L1302**: Mean platelet volume (fl), $[6.2 - 11.0]$
- **L1305**: Red blood cell count (10^{12}/L), $[3.90 - 5.50]$
- **L1306**: Neutrophils (%), $[45.0 - 75.0]$
- **L1307**: Abs. neutrophils (10^9/L), $[2.5 - 7.0]$
- **L1308**: Lymphocytes (%), $[17.0 - 55.0]$
- **L1309**: Abs. lymphocytes (10^9/L), $[0.9 - 4.5]$
- **L1310**: Monocytes (%), $[2.0 - 10.0]$

*This work was supported by NEXTCARE (COMRDI15-1-0016-2016), funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement no. 712754, the Spanish Ministry of Economy and Competitiveness under the Severo Ochoa grant SEV-2014-0425 (2015-2019), the CERCA Programme and by the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 01770)

Mireia Calvo and Raimon Jané are with Institute for Bioengineering of Catalonia (IBEC), the Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. Raimon Jané is also with Universitat Politècnica de Catalunya (UPC)-Barcelona Tech and with the Biomedical Research Networking Centre in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Barcelona, Spain {mcalvo,rjane}@ibecbarcelona.eu

Isaac Cano, Carme Hernández and Josep Roca are with Hospital Clínic de Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Centro de Investigación Biomédica en red, Enfermedades Respiratorias, University of Barcelona, E-08036, Barcelona, Spain {iscano,chernan,jroca}@clinic.cat

Vicent Ribas and Felip Miralles are with Eurecat, Technology Center of Catalonia, E-08005, Barcelona, Spain {vicent.ribas,felip.miralles}@eurecat.org

- **L1311**: Abs. monocytes (10^9/L), $[0.1 - 1.0]$
- **L1312**: Eosinophils (%), $[< 5.0]$
- **L1313**: Abs. eosinophils (10^9/L), $[< 0.5]$
- **L1314**: Hemoglobin concentration (g/L), $[120.0-170.0]$
- **L1315**: Basophils (%), $[< 2.0]$
- **L1316**: Abs. basophils (10^9/L), $[< 0.2]$
- **L1319**: Hematocrit (L/L), $[0.36 - 0.51]$
- **L1320**: Mean corpuscular volume (fl), $[80.0 - 100.0]$
- **L1321**: Mean corpuscular hemoglobin (pg), $[26.7 - 33.3]$
- **L1322**: Mean corpuscular hemoglobin concentration (g/L), $[310 - 350]$
- **L1323**: Red cell distribution width (%), $[10.5 - 17.2]$
- **L2422**: Glucose (mg/dL), $[65 - 110]$
- **L2467**: Creatinine (mg/dL), $[0.30 - 1.30]$
- **L2507**: Sodium (mEq/L), $[135 - 145]$
- **L2508**: Potassium (mEq/L), $[3.5 - 5.5]$

The study protocol was approved by the Ethical Committee for Human Research at Hospital Clínic and all patients signed a written informed consent before paticipation. Participants ages ranged between 16 and 105 (70.75 ± 15.00) years old and 62% were males. From 1951 patients being firstly admitted to the HH program, 101 were eventually transferred to regular hospitalization due to complications of heterogeneous origin (unsuccessful group). The remaining 1850 patients were discharged after a successful HH stay (successful group). Table I summarizes the baseline characteristics of both study groups, showing a significantly higher proportion of cardiovascular unsuccessful patients.

TABLE I

CLINICAL CHARACTERISTICS OF SUCCESSFUL AND UNSUCCESSFUL CASES, BASED ON THEIR NEED FOR REGULAR CARE SERVICES.

	Successful (n=1850)	**Unsuccessful (n=101)**	p**-value**
Age, *years old*	70.6 ± 15.0	72.9 ± 14.7	0.072
Male sex, *n (%)*	1153 (62.3%)	66 (65.3%)	0.613
Main diagnosis, *n (%)*			
Cardiology	196 (10.6%)	26 (25.7%)	<0.001
Respiratory	573 (31.0%)	24 (23.8%)	0.156
Oncology	146 (7.9%)	8 (7.9%)	1.000
Surgery	366 (19.8%)	15 (14.9%)	0.276
Acute	569 (30.8%)	28 (27.7%)	0.594

Values are mean ± standard deviation or number of observations (%).

B. Statistical analysis

In order to analyze the discriminant potential of blood test data and identify those variables being more susceptible of having a relevant impact on classification, we performed a first exploratory analysis where each variable was compared between successful and unsuccessful cases, by means of Mann-Whitney U non-parametric tests, setting the level of significance at $\alpha = 0.05$. Nevertheless, all variables were then taken into account for the classification approach herein proposed. Data redundancy was also evaluated by quantifying correlations between pairs of variables.

C. Class balancing

Data were divided in the training (N_{tr} = 1464, 75% randomly selected patients) and testing (N_{ts} = 487) subsets. While sampling and model training were only applied to this randomly selected training subset, the remaining 25% of data were then used for classification performance quantification. It is to note that data partitioning was performed with a stratified strategy, so as to ensure that both training and testing subsets contained a similar proportion of samples from each class as the original dataset.

Sampling methods have become standard approaches for improving classification performance in the context of imbalanced datasets [10]. They modify the training subset to create a more balanced class distribution that allows classifiers to better capture the decision boundary between majority and minority classes. The resulting (sampled) dataset represents an estimation of the original data, containing instances from the same (or similar) distribution. In this study, the following sampling techniques were overviewed:

1) Under-sampling: In random under-sampling, a subset of samples from the majority class is selected, in order to equal the number of samples coming from each target class. The main disadvantage of this strategy is that potentially relevant information from the left-out samples is lost.

2) Over-sampling: Alternatively, in random over-sampling, minority class instances are replicated. However, while avoiding information loss, it introduces the problem of overfitting. By randomly copying instances, the model might fit the training data so closely that it might not efficiently generalize to new data. Moreover, since the model can get the same samples for training and testing, model validation will be no longer independent from training and this may lead to an overestimation of the performance.

In order to overcome this limitation, Chawla et al. developed the Synthetic Minority Over-sampling TEchnique (SMOTE) [11]. With this approach, new minority instances are generated from the information contained in existing minority class samples. Another similar approach that generates new synthetic instances is the Random Over-Sampling Examples (ROSE) strategy [12], which creates new minority class samples through smoothed bootstrapping.

D. Classification performance

In order to get realistic measures of classification performance, the impact of each sampling strategy was measured and compared through cross-validation (CV), independently applying the selected sampling approach on each fold. More specifically, a 10-times repeated 4-fold CV was applied to train and test boosted logistic regression models [13].

Boosting algorithms have been reported to more suitably face class imbalance problems, since they iteratively build an ensemble of weak learners that adjust their weights to classification results. For the first weak learner, they assign equal weights to each sample, and for each subsequent learner these weights are recalculated so that higher values are assigned to previously missclassified samples. Thus, algorithms such as AdaBoost are particularly useful in this

context since they tend to give higher weights to minority class samples, often misclassified. In this study, logistic regression techniques were applied to the AdaBoost method.

Furthermore, it should be noted that, in order to equalize the contribution of all features to multivariate analysis, variables were firstly centered and scaled. Moreover, conventional classification algorithms are usually evaluated based on the percentage of observations correctly classified (accuracy). With imbalanced data, though, this performance metric might not be appropriate since minority classes hold minimum effect on overall accuracy. Thus, considering unsuccessful cases as positives, TP: true positives, FP: false positives and FN: false negatives, some interesting metrics when working with imbalanced datasets are:

- Precision (P), or positive predictive value:

$$P = \frac{TP}{TP + FP} \tag{1}$$

- Recall (R), or Sensitivity (Se):

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

- F_β measure, where β is usually taken as 1 (F_1):

$$F_\beta = (1 + \beta^2) \cdot \frac{Recall \cdot Precision}{(\beta^2 \cdot Recall) + Precision} \tag{3}$$

III. RESULTS

A. Statistical analysis

Figure 1 shows the results for correlation analysis, where it can be observed that hematocrit (L1319) is positively correlated with both hemoglobin concentration (L1314; $\rho = 0.98$) and red blood cell count (L1305; $\rho = 0.91$). The percentage of neutrophils (L306) and lymphocytes (l308), as well as their total amounts (L1300 and L1307), were found to be highly correlated ($\rho = -0.97$ and $\rho = 0.96$, respectively).

Table II summarizes the mean and standard deviation for each variable and study group, as well as the associated p-values, when Mann-Whitney U non-parametric tests were applied. Features showing statistically significant differences between groups of patients are highlighted.

Since some variables were found to be highly correlated, the application of some feature selection strategy would be advisable in the future, in order to reduce data redundancy. According to the results, the hematocrit could be omitted since it shows a very strong correlation with hemoglobin concentration, which seems to be more significant when distinguishing between groups of patients (p-value=0.030). Similarly, the percentage of neutrophils could be eliminated, in favor of lymphocytes percentage (p-value=0.040). The total amount of lymphocytes was found to be statistically significant as well (p-value=0.023). These findings concur with previous works where, on the one hand, lower levels of hemoglobin concentration indicated a poor prognosis in patients suffering from myocardial infarction [14]. On the other hand, a higher proportion of lymphocytes in the peripheral blood of patients suffering from different cancers,

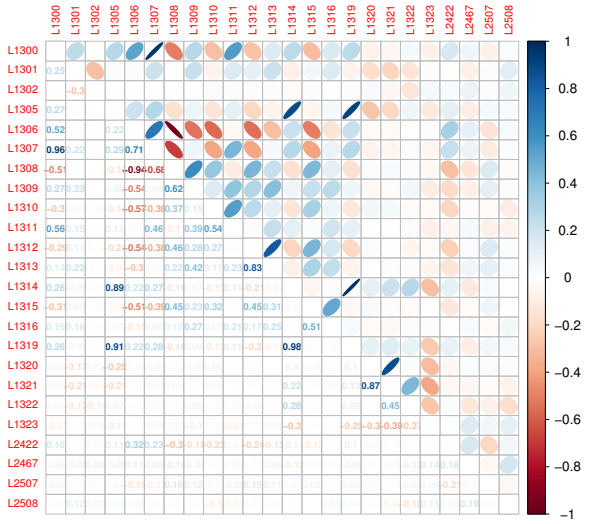

Fig. 1. Graphical and numerical representation of correlation results, for variables under study (see section II.A). Positive and negative correlations are, respectively, represented in blue and red. Numerical values are proportional to color intensity and inversely proportional to ellipse widths.

TABLE II

MEAN AND STANDARD DEVIATION FOR EACH VARIABLE (SEE SECTION II.A) AND STUDY GROUP, AND ASSOCIATED P-VALUES. *: P<0.05, FOR MANN-WHITNEY U NON-PARAMETRIC TEST.

	Successful (n=1850)	**Unsuccessful (n=101)**	***p*-value**
L1300	9.99 ± 7.48	10.09 ± 5.75	0.727
L1301	242.95 ± 107.82	266.40 ± 146.06	0.245
L1302	8.32 ± 1.06	8.54 ± 1.28	0.194
L1305	4.10 ± 0.68	3.98 ± 0.78	0.117
L1306	74.28 ± 13.24	76.01 ± 10.92	0.278
L1307	7.54 ± 4.21	8.00 ± 5.20	0.859
L1308	**15.80 ± 10.20**	**13.60 ± 7.58**	**0.040***
L1309	**1.53 ± 5.02**	**1.14 ± 0.54**	**0.023***
L1310	6.76 ± 3.91	6.80 ± 3.51	0.842
L1311	0.61 ± 0.35	0.66 ± 0.59	0.998
L1312	1.97 ± 2.18	2.26 ± 2.64	0.512
L1313	0.17 ± 0.20	0.19 ± 0.21	0.317
L1314	**121.69 ± 20.16**	**117.13 ± 21.00**	**0.030***
L1315	0.38 ± 0.86	0.38 ± 0.30	0.353
L1316	0.03 ± 0.28	0.02 ± 0.04	0.934
L1319	0.38 ± 0.06	0.36 ± 0.06	0.054
L1320	91.86 ± 6.51	91.72 ± 6.00	0.512
L1321	29.77 ± 2.38	29.63 ± 2.17	0.252
L1322	324.03 ± 11.62	322.94 ± 11.46	0.334
L1323	**14.33 ± 1.91**	**14.96 ± 2.31**	**0.002***
L2422	137.39 ± 56.87	133.71 ± 48.39	0.595
L2467	**1.03 ± 0.48**	**1.26 ± 0.75**	**0.023***
L2507	138.67 ± 3.53	138.43 ± 3.80	0.352
L2508	4.05 ± 0.50	4.15 ± 0.59	0.108

such as gastric carninomas [15], has been proposed as an indicator of poor prognosis.

Moreover, the red cell distribution width (RDW) and creatinine showed significant differences between groups of patients. These results are also in line with previous works where a higher RDW was found to be a strong independent

predictor of morbidity and mortality in heart failure [16], and where creatinine was proposed as an indicator of bad prognosis after different cardiac surgeries [7], [17].

B. Comparison of sampling approaches

Figure 2 shows the classification performance metrics obtained for each applied sampling strategy.

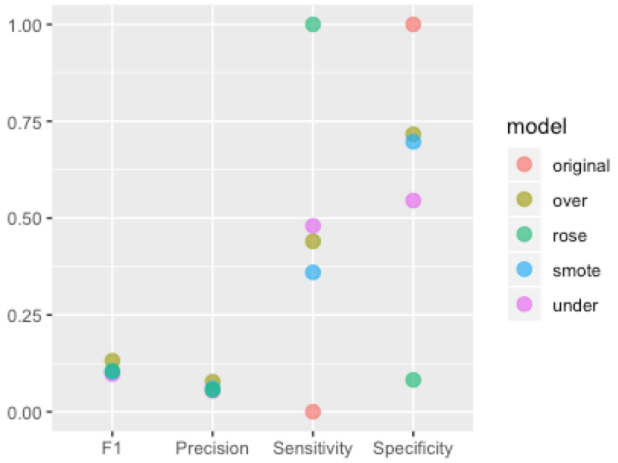

Fig. 2. Classification performance metrics for each analyzed strategy.

As expected, the original (non-sampled) model was heavily biased toward the majority class, leading to a perfect Specificity ($Sp = 1$) and a $Se = 0$, and thus not being able to measure P and F_1. The highest Se (or R) was obtained for the ROSE method, although at the expense of a very low Sp. After the original model, the SMOTE method led to the worst Se, and similar results were obtained for random under- and over-sampling. With a slightly better P, but a lower Se, random over-sampling scored the highest F_1. Nevertheless, due to class imbalance, all approaches led to very low P and thus F_1 values, since TP were gained, mostly when applying the ROSE method, at the expense of adding a large amount of FP.

IV. CONCLUSIONS

In this study, several sampling strategies were overviewed in order to face class imbalance, on a dataset containing the blood test information of 1951 patients having been admitted to a particular home-based hospitalization (HH) program.

After a first exploratory analysis, some variables, such as hemoglobin concentration, lymphocytes and creatinine, were found to unmask statistically significant differences between patients undergoing successful and unsuccessful HH stays. According to the results, blood test data is presented as a potential complementary instrument to health conditions, risk factors and socio-demographic information [18], in order to identify those patients that, if proposed for HH, should be given special attention to avoid eventual regular hospital admissions. However, class imbalance complicates the use of these data for the identification of unsuccessful cases. Among the analyzed approaches, over-sampling strategies seemed to lead to better results. The highest sensitivity was obtained with ROSE and, according to F_1, although SMOTE and random over-sampling led to similar results, the latter performed slightly better in this particular scenario.

This study presents an overview of the main sampling techniques developed to face class imbalance. In order to obtain better classification results, further improvements in the applied machine-learning approach should be performed in the future. Since significant correlations were noted among variables, a feature selection step, previous to sampling, would be advisable so as to minimize data redundancy. Moreover, although an apparently appropriate boosted model was implemented, different approaches should be explored to find the best suited model configuration. Multilevel predictive modelling including clinical, lifestyle and population data should also be considered so as to enhance the predictive performance of the proposed models.

REFERENCES

[1] S. Shepperd, S. Iliffe, "Hospital at home versus in-patient hospital care," *Cochrane Database Syst Rev*, vol. 3, 2005.

[2] K. Chevreul, *et al.*, "The development of hospital care at home: an investigation of Australian, British and Canadian experiences," *Issues Health Econ*, vol. 91, 2004.

[3] B. Leff, *et al.*, "Hospital at home: feasibility and outcomes of a program to provide hospital-level care at home for acutely ill older patients," *Ann Intern Med*, vol. 143, pp. 798–808, 2005.

[4] M. A. Mas, S. Santaeugenia, "Hospital-at-home in older patients: a scoping review on opportunities of developing comprehensive geriatric assessment based services," *Rev Esp Geriatr Gerontol*, vol. 50, no. 1, pp. 26–34, 2015.

[5] P. Tralongo, *et al.*, "Cancer patient-centered home care: a new model for health care in oncology," *Ther Clin Risk Manag*, vol. 7, pp. 387–392, 2011.

[6] S. Shepperd, *et al.*, "Admission avoidance hospital at home," *Cochrane Database Syst Rev*, vol. 9, Cd007491, 2016.

[7] F. Ryckwaert, *et al.*, "Incidence, risk factors, and prognosis of a moderate increase in plasma creatinine early after cardiac surgery," *Crit Care Med*, vol. 30, no.7, pp. 1495–1498, 2002.

[8] T. Yamanaka, *et al.*, "The baseline ratio of neutrophils to lymphocytes is associated with patient prognosis in advanced gastric cancer," *Oncology*, vol. 73, no. 3, pp. 215–220, 2007.

[9] J. J. González-Ferrer, *et al.*, "Influence of hemoglobin level on in-hospital prognosis in patients with acute coronary syndrome," *Rev Esp de Cardiol (English Edition)*, vol. 61, no. 9, pp. 945–952, 2008.

[10] N. V. Chawla, *et al.*, "Automatically countering imbalance and its empirical relationship to cost," *Data Min Knowl Discov*, vol. 17, no. 2, pp. 225-252, 2008.

[11] N. V. Chawla, *et al.*, "SMOTE: Synthetic minority over-sampling technique," *J Artif Intell Res*, vol. 16, pp. 321–357, 2002.

[12] G. Menardi, N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Min Knowl Discov*, vol. 28, no. 1, pp. 92-122, 2014.

[13] J. Friedman, *et al.*, "Additive logistic regression: a statistical view of boosting," *Ann Stat*, vol. 28, no. 2, pp. 337–407, 2000.

[14] M. M. Vis, *et al.*, "Prognostic value of admission hemoglobin levels in ST-segment elevation myocardial infarction patients presenting with cardiogenic shock," *Am J Cardiol*, vol. 99, no.9, pp. 1201–1202, 2007.

[15] J. Akagi, H. Baba, "Prognostic value of CD57+ T lymphocytes in the peripheral blood of patients with advanced gastric cancer," *Int J Clin Oncol*, vol. 13, no.6, pp. 528–535, 2008.

[16] G. M. Felker, *et al.*, "Red cell distribution width as a novel prognostic marker in heart failure: data from the CHARM Program and the Duke Databank," *J Am Coll Cardiol*, vol. 50, no.1, pp.40–47, 2007.

[17] C. Ellenberger, *et al.*, "Incidence, risk factors and prognosis of changes in serum creatinine early after aortic abdominal surgery," *J Intensive Care Med*, vol. 32, no. 11, pp. 1808–1816, 2006.

[18] C. Hernández, *et al.*, "Implementation of Home Hospitalization and Early Discharge as an Integrated Care Service: A Ten Years Pragmatic Assessment," *Int J Integr Care*, vol. 18, 2018.