ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII[th] SISMEC Congress - Vol. 20 Suppl. 1

Random Forest Regression for Predicting Healthcare Costs using Administrative Databases from a Health Protection Agency in Northern Italy

Sala Isabella[1,2], Conti Sara[3], Antonazzo Ippazio Cosimo[4], Rozza Davide[4], Losa Lorenzo[4], Ferrara Pietro[4], Fornari Carla[4], Crotti Giacomo[5], Ciampichini Roberta[5], Sampietro Giuseppe[5], Zucchi Alberto[5], Bagnardi Vincenzo[1], Mantovani Lorenzo Giovanni[4]

(1) Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy
(2) Department of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy
(3) Occupational Health Service, University of Milano-Bicocca, Milan, Italy
(4) Research Centre on Public Health, University of Milano-Bicocca, Monza, Italy
(5) Epidemiology Unit, Bergamo Health Protection Agency, Bergamo, Italy

CORRESPONDING AUTHOR: *Sala Isabella, i.sala@campus.unimib.it*

INTRODUCTION

Longer life expectancies and increasing prevalence of chronic diseases drive up demand for healthcare services and related costs. In Italy, 32% of people aged 65 and over, and 48% of those over 85, have major chronic conditions and multimorbidity [1]. In 2019, individuals aged 65 and over accounted for 46% of hospital admissions and 60% of pharmaceutical expenditures, highlighting the significant burden of aging on the healthcare system [2]. In terms of costs, population's segments with high prevalence of chronic conditions account for a large portion of healthcare spending [3,4,5]. Accurate predictions of future costs for the whole population and for key segments is crucial for healthcare planning.

AIMS

To predict yearly direct healthcare costs based on data of past National Health Service (NHS) resources utilization for the whole population and for high impacting segments. As a motivating example, we applied our approach to the dialysis patients' segment.

METHODS

Using administrative healthcare databases, we traced NHS resource utilization (i.e., access to inpatient and outpatient services, drug dispensations) and associated costs for each individual aged ≥18 assisted by the Health Protection Agency of Bergamo (Northern Italy) between 2011 and 2023. We analyzed total cost (TC) as the sum of all services and dispensations costs, total scheduled cost (TSC) as the sum of scheduled inpatient visits, all outpatient visits and dispensations costs, and scheduled services cost (SSC) as the sum of scheduled inpatient visits and all outpatient visits costs. In the present abstract we focused on TC prediction.

We used a supervised machine learning approach, namely random forest (RF) algorithm with 500 trees, to address the prediction problem [6,7]. We trained the algorithm on the 70% of individuals' data from 2011 to 2015 (n=815,553) with their TC in 2016 as outcome. The 373 input variables included demographic features (such as age and sex) and NHS utilization data over the 4-years period 2011-2014 and in 2015 alone, in order to assess if 2016 cost was more associated with subjects' behavior over the preceding year or with their historic behavior. As test sets, we used the remaining 30% of the dataset (hereafter 2011-16 set) and the subsequent years' datasets (2012-17, 2013-18, 2014-19, 2015-20, 2016-21, 2017-22, and 2018-23 sets). We considered variable importance, measured as the percent increase in mean squared error (MSE) when a given variable is permuted, as a measure of each predictor's impact on the outcome.

For each test set, actual and predicted TCs for the whole population were calculated as the sum of all individuals' actual and predicted TCs, respectively. The ratio of the difference between predicted and actual population TCs to actual population TCs was used as measure of the prediction error (PE). PE=0% indicates a perfect prediction, PE >0% or <0% suggests overestimation or underestimation of the actual TC.

DOI: 10.54103/2282-0930/29504

Finally, we defined dialysis patients as those who had at least one access to outpatient dialysis services. For this segment, we calculated the mean and sum of predicted and actual TCs, and PE. Also, we derived a variability interval for the mean predicted TC based on the 2.5 and 97.5 quantiles of the distribution of the mean TCs predicted by each tree for subjects included in the segment.

RESULTS

The mean actual annual population TC in the period from 2011 to 2023 was €1,023,636,867 (range: 944,632,707 – 1,111,657,382). High-cost subjects (>€15,000 yearly), accounting for less than 1% of the annual population, absorbed more than 27% of annual TC.

Top 3 most important variables in the RF were the number of outpatient accesses to dialysis over the preceding year, and the frequency of laboratory tests and outpatient services over the 4 preceding years.

Figure 1 shows the PEs calculated across all test sets, overall and in the dialysis patients' segment. Overall, PEs ranged from -3.1 to -1.9 across 2011-16 to 2014-19 sets (for 2014-19 set, actual annual population TC: €1,031,200,509; predicted annual population TC: €1,011,869,922), and widely increased from 2015-20 (range from -6.9 to 8.7; for 2015-20 set, actual annual population TC: €944,632,707; predicted annual population TC: €1,026,878,752)

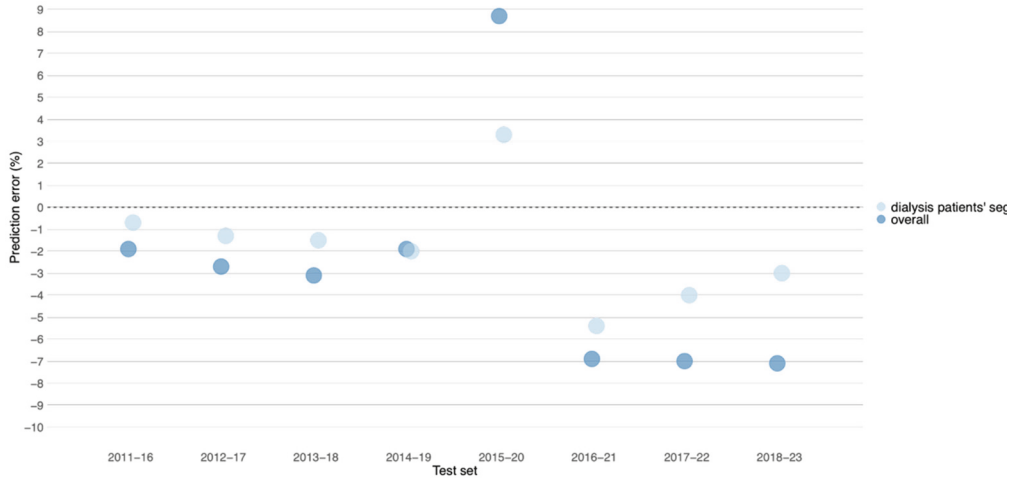
For the dialysis patients' segment, the lowest PE (-0.7%) was observed in the 2011-16 set (actual mean TC: €38,536; predicted mean TC [variability interval]: €38,259 [35,542 – 41,112]), while the highest was -5.4% in the 2016-21 set (actual mean TC: €38,883; predicted mean TC [variability interval]: €36,785 [33,967 – 39,342]).

CONCLUSIONS

Using a machine learning approach, we predicted healthcare TCs based on individual data of past utilization of NHS for the whole population and a high impacting segment. Predictions based on the algorithm trained on data from 2011 to 2015 were consistent until 2019, understandable given the COVID-19 pandemic in 2020. Results highlight the pandemic's impact on the model performance, leading to overestimation of the actual TC in 2020 and underestimations thereafter. Future steps include the identification of key segments and the update of the training algorithm on the subsequent years' datasets. This is a useful tool to assist HPA in resource allocation, e.g. as an integration to the monitoring of chronic diseases in the population.

REFERENCES

1. Istituto Nazionale di Statistica. Rapporto annuale 2024. La situazione del Paese. ISTAT (2024).
2. Istituto Nazionale di Statistica. Le Condizioni di Salute Della Popolazione Anziana in Italia, Anno 2019. ISTAT (2021). Available online: https://www.istat.it/it/files/2021/07/Report-anziani-2019.pdf [Accessed: May 2, 2025].
3. Bellini I., Barletta V., Profili F., et al., Identifying High-Cost, High-Risk Patients Using Administrative Databases in Tuscany, Italy. BioMed Research International, 2017(1), 9569348.
4. Ricciardi W, Tarricone R., The evolution of the Italian National Health Service. The Lancet., 2021; 398(10317):2193–206.
5. Blumenthal D, Chernof B, Fulmer T, et al., Caring for high-need, high-cost patients - an urgent priority. N Engl J Med., 2016; 375(10):909–11.
6. Breiman L., Random forests. Machine learning, 2001; 45, 5-32.
7. Breiman L., Cutler A., Liaw A., et al., Breiman and Cutler's random forests for classification and regression. R package version, 3(3). 2018.



	2011–16	2012–17	2013–18	2014–19	2015–20	2016–21	2017–22	2018–23
N	242	1,191	1,227	703	1,239	1,181	1,278	1,297
Mean PTC	€38,259	€38,568	€37,497	€37,877	€37,316	€36,785	€37,015	€36,242
riability interval for mean PTC	(35,542 – 41,112)	(36,137 – 40,978)	(34,783 – 40,167)	(35,076 – 40,267)	(34,344 – 40,100)	(33,967 – 39,342)	(34,305 – 39,828)	(33,266 – 38,927)
Mean ATC	€38,536	€39,067	€38,082	€38,668	€36,134	€38,883	€38,547	€37,363
N	244,666	819,264	821,286	823,146	824,801	821,859	824,076	826,635
PTC	€289,421,614	€977,616,321	€993,861,272	€1,011,869,922	€1,026,878,752	€932,942,579	€1,009,175,682	€1,032,554,950
ATC	€294,927,249	€1,003,146,952	€1,026,029,933	€1,031,200,509	€944,632,707	€1,002,438,689	€1,085,121,065	€1,111,657,382

Figure 1. Prediction errors of total cost, overall (blue dots) and for the dialysis patients' segment (light blue dots) across all test sets. Abbreviations: ACT=actual total cost, PTC=predicted total cost.

DOI: 10.54103/2282-0930/29504