


RESEARCH ARTICLE

Comparing risk adjustment estimation methods under data availability constraints

Marica Iommi¹  | Savannah Bergquist² | Gianluca Fiorentini³ | Francesco Paolucci^{4,5}

¹Advanced School for Health Policy,
University of Bologna, Bologna, Italy

²Haas School of Business, University of
California, Berkeley, California, USA

³Department of Economics, University of
Bologna, Bologna, Italy

⁴Newcastle Business School, University of
Newcastle, Newcastle, Australia

⁵School of Economics and Management,
University of Bologna, Bologna, Italy

Correspondence

Marica Iommi, Advanced School for Health
Policy, University of Bologna, Bologna,
Italy.

Email: marica.iommi2@unibo.it

Funding information

Agenzia Sanitaria e Sociale Regionale,
Regione Emilia-Romagna; Università degli
Studi di Bologna

Open Access Funding provided by
Università degli Studi di Bologna within the
CRUI-CARE Agreement.

Abstract

The Italian National Healthcare Service relies on per capita allocation for healthcare funds, despite having a highly detailed and wide range of data to potentially build a complex risk-adjustment formula. However, heterogeneity in data availability limits the development of a national model. This paper implements and evaluates machine learning (ML) and standard risk-adjustment models on different data scenarios that a Region or Country may face, to optimize information with the most predictive model. We show that ML achieves a small but generally statistically insignificant improvement of adjusted R^2 and mean squared error with fine data granularity compared to linear regression, while in coarse granularity and poor range of variables scenario no differences were observed. The advantage of ML algorithms is greater in the coarse granularity and fair/rich range of variables set and limited with fine granularity scenarios. The inclusion of detailed morbidity- and pharmacy-based adjusters generally increases fit, although the trade-off of creating adverse economic incentives must be considered.

KEYWORDS

data granularity, formula funding, health expenditure, machine learning, risk-adjustment

1 | INTRODUCTION

In middle- and high-income countries, health expenditure has steadily increased over time, making its containment a major issue for governments: policymakers must control healthcare resources to efficiently address population aging and public budget constraints. Risk-adjustment schemes are one such tool for managing the efficiency and fairness of healthcare spending (Cid et al., 2016). The primary statistical purpose of risk-adjustment models, both in a health plan payment or provider reimbursement, is to accurately predict individual healthcare costs. When designing a risk-adjustment program, policymakers should aim to optimize the choice of statistical methods based on their data availability in the short-medium term, pending a longer-term investment to expand the range of variables and level of detail.

The performance of a risk-adjustment algorithm depends not only on the model or functional form, but also on the underlying data. In a setting with decentralized data administration, differences in data collection – and thus in the data available for prediction – stem from several sources, including: the accessibility of distinct databases (inpatient, pharmaceutical, outpatient services), the range of data elements collected, the ability to merge information at an individual level (due to privacy

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. Health Economics published by John Wiley & Sons Ltd.

regulations), and different quality standards or level of detail in data collection (rudimentary vs. advanced information systems) according to the local context (we refer to the level of detail or aggregation as the “granularity” of the data).

The Italian National Healthcare Service (NHS) is an example of a healthcare system with substantial heterogeneity in data collection. With its decentralized units (21 health regions) all subject to the same national legislation regarding data collection duties, the Italian NHS has the potential for using a highly detailed and wide-ranging set of information to predict individual health expenditure. However, data collection takes place at the local level, where there are differing capacities to combine healthcare administrative databases, heterogeneous information systems, and variability in data granularity and quality control (Skrami et al., 2019). Thus, despite this potential richness of individual health information, the allocation of Italian NHS resources to Regions is still based on age-weighted capitation. Per capita allocation using synthetic indices methods or actuarial cells formulas is common among national health systems (e.g., Spain, Denmark, Scotland, Portugal) to distribute resources in a large geographical area for its simplicity of construction and low cost information (Cid et al., 2016). However, person-based formulae are increasingly used as an alternative method (e.g., United Kingdom), and they offer considerable scope for ensuring equitable allocations of health budgets to geographical areas (Radinmanesh et al., 2021; Smith, 2006).

In this paper, we develop a framework for designing an effective risk-adjustment algorithm to predict total healthcare expenditure by using routinely available, person-based Italian data on health needs. We explore different data scenarios by examining how the predictive performance of conventional and machine learning (ML) risk-adjustment estimation varies across by data granularity and range of variables settings.

Our primary contribution is to develop general guidelines for policymakers who have various scenarios of data granularity and/or range of variables and to provide a set of applicable statistical tools. For countries with national health systems this study may serve as a basis for budget setting using person-based regression models, which are more predictive than synthetic indices methods or actuarial cells formulas because they capture more variability in patient needs (Ellis, 2007). Countries with data constraints may use our framework to optimize the information available to improve efficiency and fairness in allocating resources or to reduce incentives for risk-selection (Henríquez et al., 2020). Countries with less limiting data constraints can implement more sophisticated ML techniques to improve predictive performances and fit efficiency (Geruso & McGuire, 2016; Kronick & Welch, 2014; Rose, 2016).

2 | EMPIRICAL FRAMEWORK

In Table 1 we present six typical data scenarios to predict total healthcare expenditure, which varies in two dimensions: range of variables (the number of health databases), and data granularity (the detail of the information in each database). These scenarios are based on real-world implementations of risk-adjustment programs and data collection; it is common to begin with age-sex cells, then add in hospitalization information, and finally pharmaceutical data. Similarly, a coarser level of detail may be more amenable to basic data collection and simpler prediction methods. The Italian health information system is theoretically located in the cell at the bottom right (fine data/rich range of variables), but as long as the reliability and consistency of the administrative databases is uncertain, the first cell at the top left (coarse data/poor range of variables) could be the first plausible scenario in the transition from synthetic indices methods to more predictive models.

We use “range of variables” to refer to the data availability with respect to different health areas, and with “data granularity” we refer to the detail available. Age and sex are optimal exogenous risk-adjusters because they cannot be influenced by health providers' actions. However, they do not accurately reflect the epidemiological variability and health needs among geographical areas (Atella et al., 2018). Since the late 1990s, the United States (US) and many European health systems expanded the range of variables to reduce risk selection with increasingly sophisticated risk-adjustment formulas, including socio-economic, pharmacy- and diagnosis-based indicators (Ellis et al., 2018). Diagnosis information may be endogenous to some extent but is generally far more predictive than age and sex alone. In the UK, factors capturing demography, morbidity, deprivation, and unavoidable cost of providing services in different areas have been included in a complex person-based formula to improve the allocation of the “fair share” for each area (Dixon et al., 2011).

Early studies in the US demonstrated the inclusion of diagnosis cost groups (DCGs) improved the R^2 of the demographic model (age-sex data) by 40% points in a concurrent model and by 10% points in a prospective model (Ellis et al., 1996; Pope et al., 2000). A study for allocating commissioning funds to general practices in England, showed that from a model with age and sex only, adding diagnostic-related variables improved predictive power from 3.7% to 12.6% at individual level (Dixon et al., 2011).

In 2002, the Dutch health system was the first to expand the range of variables with pharmaceutical data through the Pharmacy-based Cost Group (PCG), a risk-adjuster based on pharmaceutical use, which increased the R^2 from 4% in the

TABLE 1 Data setting scenarios to predict total health care expenditure

| | Poor range of variables | Fair range of variables | Rich range of variables |
|--|--|--|---|
| | Demographic (DEM) | DEM + hospital discharge records (HDR) | DEM + HDR + pharmacy database (PD) |
| Coarse granularity (sparse detail) | <i>Total health care costs = f(sex * age groups)</i> | <i>Total health care costs = f(sex * age groups + n. of hospitalizations)</i> | <i>Total health care costs = f(sex * age groups + n. of hospitalizations + n. of drug prescriptions)</i> |
| | Age groups: 0, 1–4, 5-year classes up to 90, 90+ | HDR: N. of hospitalizations | HDR: N. of hospitalizations PD: N. of prescriptions |
| Fine granularity (high detail) | <i>Total health care costs = f(sex * age groups + citizenship + degree of urbanization + income)</i> | <i>Total health care costs = f(sex * age groups + citizenship + municipality of residence + income + group of ICD-9 codes)</i> | <i>Total health care costs = f(sex * age groups + citizenship + municipality of residence + income + group of ICD-9 codes + group of ATC codes)</i> |
| | Age groups: 0, 1–4, 5-year classes up to 90, 90+ | HDR: ICD-9-CM codes which can be classified into diagnosis-related groups ^a | HDR: ICD-9-CM codes; may be classified into DCGs PD: ATC codes; may be classified into PCGs ^b |
| | Citizenship: Italian (yes/no) | | |
| | Degree of urbanization: Urban, peri-urban, rural | | |
| | Income: low, medium-low, medium high, high | | |

Abbreviations: ESRD, End-Stage Renal Disease; HCC, Hierarchical Condition Category; HDR, hospital discharge records; PD, pharmacy database; RxHCC, Prescription Drug Hierarchical Condition Category.

^aFor example, possible coding classification systems include the Clinical Classifications Software for ICD-9-CM (<https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>) or the 2015 Risk Adjustment model software HCC, RxHCC, ESRD of Centers for Medicare & Medicaid Services (<https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors>) or the DCG.

^bFor example, Pharmacy Cost-Group (PCG).

demographic model to almost 9% (Lamers, 1999). In 2017, the sophisticated prospective Dutch risk-adjustment model reached an R^2 of almost 32% (van Kleef et al., 2018).

Other countries with competitive insurance markets have introduced morbidity indicators in their risk adjustment modeling, albeit with some limitations. Switzerland introduced PCGs into risk-adjustment modeling after a study found that the inclusion of pharmaceutical indicators in the Swiss prospective formula increases the R^2 by 13% points (Schmid & Beck, 2016), while diagnostic information is still incorporated as a dummy variable (yes/no prior hospitalizations). Germany uses pharmaceutical data to validate or impute missing diagnoses, while the Netherlands does not include outpatient diagnoses in the risk-adjustment systems (Bauhoff et al., 2017). Most US risk-adjustment systems do not use pharmaceutical data as predictors for health care expenditure (except Marketplaces), although Wagner et al. showed the DxCG system (supported by Verisk, a commonly used proprietary risk adjustment systems for cost data), which computes risk score using pharmacy and hospital data, increased the R^2 estimates by 8 to 20% points compared to the base models without pharmacy data (Wagner et al., 2016).

Despite the development of indicators based on diagnosis and pharmaceutical data, which have substantially improved the predictive performance of risk-adjustment algorithms over the past 2 decades, data constraints are still present in many countries. In Italy, approximately 35% (outpatient service and half of hospital expenditure) of the Italian National Health Fund is distributed to Regions based on an age-weighting system, while the remainder is allocated based on the unweighted resident population (Ferre et al., 2014). Ireland's health system relies on age, gender, and level of coverage as risk adjusters to determine transfers under risk equalization (Armstrong, 2018); Australian and Chilean risk equalization formulas are based on age, gender, and location (Paolucci et al., 2018; Velasco et al., 2018).

3 | NEW RISK-ADJUSTMENT ESTIMATION TECHNIQUES

Ordinary least squares (OLS) is the tool most commonly used in practice to predict healthcare spending (Ellis et al., 2018; Iezzoni, 2012). Recent literature has explored deploying ML methods, particularly in health systems with fine data and rich range of variables. The exponential growth of the size and complexity of electronic health information and the advancement

in computational capacity has shifted the attention of health statisticians and econometricians towards newer techniques (Rose, 2016). These techniques allow for the automated investigation of all possible covariates available and to leverage the complex structure of population-level health data (Kan et al., 2019).

Rose (2016) compared OLS with 8 ML algorithms (lasso, ridge, elastic net, neural net, single tree, random forests (RF), and a discrete and weighted super learner (SL)) using the inputs of the official US individual Marketplace risk-adjustment formula, including age, sex, geographic area, 5 inpatient diagnosis categories, and 74 Hierarchical Condition Category (HCC) variables, to predict total annual expenditures in a prospective formula. The results highlighted the minor improvement in performance of the SL with respect to cross-validated R^2 , but also showed that reducing the number of covariates to 10 only marginally reduced the predictive performance, providing preliminary evidence that few variables could yield effective plan payment risk-adjustment. McGuire, Zink, and Rose (McGuire et al., 2020) examined payment system fit with a concurrent formula in the setting of the US individual Marketplaces. The study found a penalized regression with age and sex cells and 30 HCCs performed similarly to the baseline linear regression with age and sex cells and 90+ HCC-based variables, while pharmaceutical covariates did not improve fit over the diagnosis-based risk adjusters. In recent work, Kan et al. (2019) compared OLS and penalized linear regressions for predicting Medicare Advantage health care expenditure (patients aged ≥ 65 years). The authors emphasized the advantages of penalized regressions, particularly the lasso, which yielded improvements over OLS in a single year of data and a 4-year pooled-data setting.

In non-US settings, exploration of interaction terms has been common. Buchner et al. employed a regression tree analysis to improve the German risk-adjustment formula with significant interactions between variables (Buchner et al., 2017). The inclusion of morbidity interaction terms showed a marginal improvement of R^2 , with an insubstantial loss in accuracy. Examining the 2014 Dutch risk-adjustment formula, van Veen et al. (van Veen et al., 2018) found that a similar approach improved the adjusted R^2 , but concluded it may also increase the possibility of risk-selection for some subgroups.

4 | DATA

Our study population was drawn from the 2016 administrative databases of Emilia, a Northern Italian Region with about 4.4 million beneficiaries with universal access to the Italian NHS.

The main databases used were:

1. Hospital Discharge Records: Admissions and discharge dates; primary and up to five secondary diagnoses and up to six interventions (ICD-9-CM coding system) are included. Expenditure is registered using the diagnosis-related group tariffs system since 1995 (Emilia-Romagna, 2014).
2. Outpatient Pharmaceutical Database (OPD): Contains drugs reimbursed by the NHS (prescribed by the primary care physician or a specialist, or directly dispensed by the hospital pharmacies) and details on substance name, Anatomical Therapeutic Chemical (ATC) classification system code-V.2013, brand name, date of prescription filling, number of unit doses and number of packages and prescribers.
3. Outpatient Speciality Database (OSD): Contains individual expenditures of laboratory test and specialistic visits.

The linkage among these databases was possible through a unique anonymized patient identifier. We include only individuals who were residents in 2016 of one of the 8 Local Health Authorities of Emilia-Romagna.

4.1 | Socio-demographic covariates

We included gender, age (divided into 20 classes: 0, 1-4, 5-year classes up to 90 years and a class open over 90 years), citizenship (dichotomous variable: Italian yes/no), the degree of urbanization of residence (classified using the Eurostat's Degree of Urbanization (DEGURBA) (Eurostat, 2014) classification system into sparsely populated areas, intermediate density areas, and densely populated areas) and the income derived from the exemptions (if an individual presented more than one exemption they were assigned to the lowest income class).

4.2 | Hospital covariates

The hospital covariates included the total number of hospitalization and ICD-9-CM codes, which were grouped into DCGs. This method, originally proposed by Ash and colleagues (Ash et al., 1989; Ellis & Ash, 1995), first aggregates the diagnostic codes into 78 clinically homogeneous subgroups, then further aggregates these subgroups into 9 large groups on the basis of their equivalence in terms of resource absorption. Individuals with multiple diagnoses were classified only in their most expensive DCG while individuals without any diagnoses were grouped into DCG 0. Appendix Table A contains the full list of DCGs.

4.3 | Pharmacy covariates

From the OPD, we included the total number of drug prescriptions and ATC codes, which we grouped into PCGs. Through the ATC coding system, individuals with pharmaceutical prescriptions in 1 year are marked with a medical condition, which is then used as risk adjuster for the distribution of the resources thereafter. For the construction of the PCGs, we combined the Dutch model and other approaches that use the ATC codes to identify patients with chronic condition from the pharmaceutical databases defining 36 distinct classes (Chini et al., 2011; Corrao et al., 2017; Trottmann et al., 2010; Trottmann et al., 2015). An individual is classified into a PCG if the year amount of defined daily doses (DDD) in a given pharmaceutical class was greater than 180 DDD (cancer drugs required only >15 DDD). The full list of PCGs is shown Appendix Table B.

4.4 | Dependent variable

The outcome of interest is annual total expenditure, calculated by summing up the expenditures of hospital, pharmaceutical, and outpatient care services provided to residents of Emilia-Romagna Region during 2016.

We implemented a concurrent risk-adjustment approach: health care expenditure was predicted by individual characteristics from the same year. The concurrent method, compared to the prospective, reduces the data burden by using only a single year of data and does not require a separate formula for individuals with no information from the prior year. Concurrent risk models generally yield higher predictive power than prospective models. The compromise of this higher predictiveness is a reduced incentive to control healthcare spending and a longer timeline for payment, adding uncertainty, administrative burdens, and planning challenges (Geruso & McGuire, 2016).

4.5 | Characteristics of the population

A summary of socio-demographic variables is presented in Table 2. There were slightly more women than men in the sample (52%) and the largest age groups were 50–54 (9.2%) and 45–49 (9.1%). Mean total healthcare expenditure was 834€. Women had a slightly lower mean expenditure compared to men (832 vs. 841€) but a much higher median expenditure (114 vs. 59€). Italians had higher mean expenditure than non-Italians (861 vs. 530€) and people who lived in rural areas had higher expenditure (876€) than those living in urban (858€) or peri-urban areas (796€). Low-income exempted individuals had the highest mean and median expenditure (1161 and 226€) compared with other income groups. Average spending increased with more hospitalizations or prescriptions. Similarly, individuals with DCG were more expensive (5631€) than individuals with no DCG (227€), as well for individuals with at least one PCG (1729 €) compared to those with no PCG (406 €).

5 | STATISTICAL ANALYSIS

5.1 | Model specifications

As our baseline method, we implemented an OLS regression. To account for the highly skewed distribution of healthcare spending, we also estimated generalized linear models, using the frequently applied log-link function and two distribution families (Gaussian and Gamma; Jones, 2010).

TABLE 2 Description of the study population ($n = 4,262,982$)

| | Total expenditure in 2016 | | | | | | |
|------------------------------------|---------------------------|-------|----------|----------|-----------------------|--------|-----------------------|
| | N | % Col | Mean | S. D. | 25 th perc | Median | 75 th perc |
| Observations | 4,262,982 | 100% | 834.0 | 3306.0 | 13 | 87 | 405 |
| Male | 2,054,695 | 48.2% | 838.2 | 3618.0 | 6 | 59 | 344 |
| Female | 2,208,287 | 51.8% | 829.8 | 2986.9 | 24 | 114 | 460 |
| Age groups | | | | | | | |
| 0 | 48,632 | 1.1% | 1358.1 | 4832.4 | 128 | 340 | 416 |
| 1–4 | 119,804 | 2.8% | 234.6 | 1875.0 | 11 | 30 | 82 |
| 5–9 | 157,598 | 3.7% | 167.7 | 1339.2 | 7 | 21 | 65 |
| 10–14 | 153,215 | 3.6% | 201.9 | 1448.7 | 6 | 29 | 99 |
| 15–19 | 144,999 | 3.4% | 244.1 | 1873.1 | 5 | 24 | 98 |
| 20–24 | 148,695 | 3.5% | 286.8 | 1870.2 | 3 | 22 | 91 |
| 25–29 | 184,774 | 4.3% | 322.5 | 1795.3 | 4 | 26 | 111 |
| 30–34 | 225,961 | 5.3% | 358.4 | 1730.4 | 5 | 31 | 131 |
| 35–39 | 279,342 | 6.6% | 374.3 | 1925.0 | 5 | 31 | 138 |
| 40–44 | 353,249 | 8.3% | 384.8 | 2267.9 | 2 | 32 | 141 |
| 45–49 | 389,168 | 9.1% | 424.2 | 2359.5 | 4 | 48 | 172 |
| 50–54 | 391,461 | 9.2% | 528.0 | 2883.7 | 4 | 57 | 223 |
| 55–59 | 313,161 | 7.3% | 743.2 | 3579.0 | 18 | 103 | 337 |
| 60–64 | 266,900 | 6.3% | 1008.0 | 3978.9 | 40 | 172 | 497 |
| 65–69 | 273,682 | 6.4% | 1378.1 | 4351.7 | 101 | 318 | 809 |
| 70–74 | 224,178 | 5.3% | 1771.5 | 4675.0 | 173 | 459 | 1145 |
| 75–79 | 226,081 | 5.3% | 2077.4 | 5020.8 | 237 | 568 | 1414 |
| 80–84 | 173,644 | 4.1% | 2277.6 | 4929.6 | 287 | 638 | 1690 |
| 85–89 | 116,705 | 2.7% | 2325.1 | 4592.5 | 275 | 617 | 2082 |
| 90+ | 71,733 | 1.7% | 2225.8 | 4002.1 | 205 | 506 | 2713 |
| Italian citizenship | | | | | | | |
| No | 351,815 | 8.3% | 529.9 | 2761.0 | 8 | 50 | 220 |
| Yes | 3,911,167 | 91.7% | 861.2 | 3349.5 | 13 | 92 | 425 |
| Degree of urbanization | | | | | | | |
| Urban | 1,548,231 | 36.3% | 857.8 | 3418.7 | 11 | 84 | 406 |
| Peri-urban | 1,896,501 | 44.5% | 796.1 | 3191.1 | 13 | 83 | 385 |
| Rural | 818,250 | 19.2% | 875.9 | 3350.3 | 17 | 102 | 450 |
| Income | | | | | | | |
| Exempt low income | 1,746,677 | 41.0% | 1161.1 | 3576.4 | 54 | 226 | 735 |
| Exempt middle-low income | 583,624 | 13.7% | 720.9 | 2767.8 | 24 | 114 | 378 |
| Exempt middle-high income | 98,832 | 2.3% | 696.8 | 2777.1 | 17 | 94 | 334 |
| Exempt high income | 397,233 | 9.3% | 797.9 | 3817.2 | 0 | 38 | 235 |
| Exempt other reason | 1,436,616 | 33.7% | 501.2 | 2994.1 | 1 | 22 | 119 |
| No. of hospitalization | | | | | | | |
| 0 | 3,777,073 | 88.6% | 226.3 | 728.4 | 9 | 61 | 245 |
| 1 | 380,626 | 8.9% | 3768.6 | 5440.2 | 1362 | 2297 | 4272 |
| 2 | 71,425 | 1.7% | 9101.3 | 8904.1 | 3974 | 6841 | 11,434 |
| 3+ | 33,858 | 0.8% | 18,174.6 | 15,304.0 | 9145 | 14,355 | 22,410 |
| No. of pharmaceutical prescription | | | | | | | |
| 0 | 931,697 | 21.9% | 205.5 | 1726.9 | 0 | 21 | 75 |

(Continues)

TABLE 2 (Continued)

| | Total expenditure in 2016 | | | | | | |
|-----------------------------|---------------------------|-------|--------|--------|-----------------------|--------|-----------------------|
| | N | % Col | Mean | S. D. | 25 th perc | Median | 75 th perc |
| 1 | 876,850 | 20.6% | 158.3 | 1244.2 | 0 | 5 | 28 |
| 2 | 326,076 | 7.6% | 350.6 | 1895.8 | 15 | 34 | 116 |
| 3+ | 2,128,359 | 49.9% | 1461.3 | 4313.9 | 115 | 322 | 904 |
| Diagnostic cost group (DCG) | | | | | | | |
| No | 3,784,039 | 88.8% | 226.7 | 730.9 | 9 | 61 | 245 |
| Yes | 478,943 | 11.2% | 5631.1 | 8194.2 | 1614 | 3033 | 6466 |
| Pharmacy cost group (PCG) | | | | | | | |
| No | 2,884,260 | 67.7% | 406.0 | 2302.7 | 5 | 34 | 131 |
| Yes | 1,378,722 | 32.3% | 1728.8 | 4639.1 | 186 | 446 | 1109 |

Abbreviations: DCG, Diagnostic cost group; PCG, Pharmacy cost group.

We deployed the following ML algorithms: penalized regressions (lasso, ridge, and elastic net penalties), a generalized additive model (GAM), RF, and, finally, we constructed a SL, an ensembling method that takes a weighted average of predicted values from underlying candidate algorithms (such as penalized regressions, GAMs, and RFs) to produce a single best prediction function. (M. J. van der Laan & Rose, 2011; Mark J. van der Laan et al., 2007). For further background on these methods and their application to risk adjustment we refer the interested reader to Rose (2016).

5.2 | Model performance

We randomly sampled 100,000 observations from our total sample. We divided this subset a training set (70%), to develop and validate the algorithms, and test set (30%) to evaluate the performance of the models (Appendix C reports summary statistics for the train and test sets). To choose the optimal tuning parameters of the penalized linear regressions, 10-fold cross-validation was used within the training set. For the RF, 500 trees were produced (node size = 250) and p/3 predictors, for each specification formula, were randomly chosen at each step of the tree-building.

Model fit was judged using the adjusted-R² and mean squared error (MSE). Predictive Ratios (PR) were computed as the ratio between mean predicted spending and mean observed spending in each observed quintile. We also calculated mean under/overcompensation, computed as the difference between mean predicted spending and the mean observed spending in each quintile. For MSE, a smaller value indicates better performance. An adjusted-R² a value closer to one indicates more explanatory power, while PR values between 0.9 and 1.1 are considered reasonable prediction accuracy (Kautter et al., 2012). The Relative Efficiency (RE) of each algorithm compared to OLS, with respect to both cross-validated MSE ($RE = cv\ MSE_{OLS}/cv\ MSE_k$, $k = \text{other algorithms}$) and R² values ($RE = cv\ R^2_k/cv\ R^2_{OLS}$, $k = \text{other algorithms}$), was evaluated. A total of 54 models were estimated to predict healthcare expenditure (based on six data scenarios and nine algorithms). Statistical analyses were performed using SAS Enterprise Guide (version 7.1) and R (version 3.6.3).

6 | RESULTS

6.1 | Evaluation of the models: Adjusted-R² and MSE

Tables 3 and 4 report the best and worst performing models compared to OLS according to adjusted-R² and MSE, respectively. In all scenarios, SL displayed the best performance, except in the coarse granularity/poor range of variables setting, in which the lasso model performed best (in both adjusted-R² and MSE measures). However, in this setting, the performance difference between models was trivial, such that the RE was 1 across all models. Focusing on the adjusted-R², the gain in relative efficiency compared to OLS was higher in the coarse granularity/rich range of variables setting (SL adj.-R² = 49.9%) and in the fine granularity/poor range of variables (SL adj.-R² = 4.6%), but in both these settings there was no gain in RE. In the fine granularity/fair and rich range of variables scenarios, the SL RE was 1.0, with an adjusted-R² of 43.6% and 45.1% when DCG and DCG+PCG were added, respectively. We obtained inconsistent values (the predictions were so poor that the adjusted-R²

TABLE 3 Adjusted-R² of the models compared to OLS

| | Poor range of variables | | | Fair range of variables | | | Rich range of variables | | |
|--------------------|-------------------------|---------------------|----------------|-------------------------|---------------------|----------------|-------------------------|---------------------|------------|
| | Models ranking | Adj. R ² | RE | Models ranking | Adj. R ² | RE | Models ranking | Adj. R ² | RE |
| Coarse granularity | <i>Lasso</i> | 4.3% | 1.0 | <i>SL</i> | 48.8% | 1.0 | <i>SL</i> | 49.9% | 1.0 |
| | <i>ElNet</i> | 4.3% | 1.0 | <i>GAM</i> | 48.0% | 1.0 | <i>RF</i> | 49.1% | 1.0 |
| | <i>RF</i> | 4.3% | 1.0 | <i>OLS</i> | 47.6% | 1.0 | <i>GAM</i> | 48.9% | 1.0 |
| | <i>SL</i> | 4.3% | 1.0 | <i>Lasso</i> | 47.6% | 1.0 | <i>OLS</i> | 48.5% | 1.0 |
| | <i>GLM l-g</i> | 4.3% | 1.0 | <i>ElNet</i> | 47.6% | 1.0 | <i>Lasso</i> | 48.5% | 1.0 |
| | <i>GLM l-n</i> | 4.3% | 1.0 | <i>RF</i> | 47.5% | 1.0 | <i>ElNet</i> | 48.5% | 1.0 |
| | <i>OLS</i> | 4.3% | 1.0 | <i>Ridge</i> | 47.3% | 1.0 | <i>Ridge</i> | 48.2% | 1.0 |
| | <i>GAM</i> | 4.3% | 1.0 | <i>GLM l-n</i> | Inc. value | - | <i>GLM l-n</i> | Inc. value | - |
| | <i>Ridge</i> | 4.3% | 1.0 | <i>GLM l-g</i> | Inc. value | - | <i>GLM l-g</i> | Inc. value | - |
| Fine granularity | Models ranking | Adj. R ² | RE | Models ranking | Adj. R ² | RE | Models ranking | Adj. R ² | RE |
| | <i>SL</i> | 4.6% | 1.0 | <i>SL</i> | 43.6% | 1.0 | <i>SL</i> | 45.1% | 1.0 |
| | <i>GLM l-n</i> | 4.6% | 1.0 | <i>GLM l-n</i> | 43.2% | 1.0 | <i>Lasso</i> | 44.9% | 1.0 |
| | <i>RF</i> | 4.5% | 1.0 | <i>Lasso</i> | 43.2% | 1.0 | <i>ElNet</i> | 44.9% | 1.0 |
| | <i>Lasso</i> | 4.5% | 1.0 | <i>ElNet</i> | 43.2% | 1.0 | <i>OLS</i> | 44.9% | 1.0 |
| | <i>ElNet</i> | 4.5% | 1.0 | <i>OLS</i> | 43.2% | 1.0 | <i>GAM</i> | 44.9% | 1.0 |
| | <i>OLS</i> | 4.5% | 1.0 | <i>GAM</i> | 43.2% | 1.0 | <i>Ridge</i> | 44.8% | 1.0 |
| | <i>GAM</i> | 4.5% | 1.0 | <i>Ridge</i> | 43.0% | 1.0 | <i>GLM l-n</i> | 44.6% | 1.0 |
| | <i>Ridge</i> | 4.5% | 1.0 | <i>RF</i> | 42.8% | 1.0 | <i>RF</i> | 44.5% | 1.0 |
| <i>GLM l-g</i> | 4.5% | 1.0 | <i>GLM l-g</i> | 6.6% | 0.2 | <i>GLM l-g</i> | Inc. value | - | |

Abbreviations: GAM, generalized additive model; MSE, mean squared error; OLS, Ordinary least squares; RF, random forest; SL, super learner.

were negative and the MSEs were extremely high) for the GLMs, denoting their poor performances. The lowest adjusted-R² of OLS was 4.3% in the coarse granularity/poor range of variables data while the highest was 48.5%, obtained in the coarse granularity/rich range of variables data.

Concerning MSE, in the coarse granularity/poor range of variables and fine granularity/fair and rich range of variables scenarios, there were no gains in RE of using SL compared to OLS. ML technique produced slight gain in the coarse granularity/fair range and rich range of variables scenarios, where SL reduced the MSE by 2.1% (MSE = 5,705,097; RE = 1.02) and by 2.6% (MSE = 5,583,558; RE = 1.03). The GLMs yielded extremely high MSE values and the worst performance in many scenarios. The worst MSE of OLS was in the coarse granularity/poor range of variables data (MSE = 10,657,875) while the best was in the coarse granularity/rich range of variables data (MSE = 5,734,502).

We carried out further analyses, in which we compared the performance of the ML models, estimated on a subsample of the population as above, with that of the OLS model, estimated on the entire population of 4,262,982 individuals (“OLS total”). The results are presented in the appendix D and E. In the coarse and fine granularity/poor range of variables scenarios, OLS total outperformed the best ML algorithm for both adjusted-R² and MSE. Concerning MSE, the gap between the best ML model and the OLS total model narrows, until it reverses the sign in the fine granularity/rich range of variables scenario.

6.2 | Under/overcompensation and predictive ratios

Figure 1 shows the mean under/overcompensation by observed spending quintile. Complete tables for each data scenario of mean under/overcompensation and PRs are reported in Appendix F and G. In the coarse granularity/poor range of variables scenario all models highly overcompensate the expenditure of the individuals in the 80th percentile group (best: ridge = 822.8; worst: OLS = 826.5) as well as in the fine granularity/poor range of variables (best: SL = 859.7; worst: OLS = 868.9). In the 99th percentile all models highly undercompensate. In contrast, the 99th percentile group is (in a lesser extent) overcompensated in the fair and rich range of variables settings, for both coarse and fine granularity. Finally, in the coarse granularity/rich range of variables setting we also see the lowest under/overcompensation relative to the other scenarios for the 20th percentile.

TABLE 4 Average cross-validated MSE of the models compared to OLS

| | Poor range of variables | | | Fair range of variables | | | Rich range of variables | | |
|--------------------|-------------------------|-------------------|------------|-------------------------|------------------|------------|-------------------------|------------------|------------|
| | Models ranking | MSE | RE | Models ranking | MSE | RE | Models ranking | MSE | RE |
| Coarse granularity | <i>Lasso</i> | 10,657,793 | 1.0 | <i>SL</i> | 5,705,097 | 1.0 | <i>SL</i> | 5,583,558 | 1.0 |
| | <i>ElNet</i> | 10,657,794 | 1.0 | <i>GAM</i> | 5,791,326 | 1.0 | <i>RF</i> | 5,663,559 | 1.0 |
| | <i>RF</i> | 10,657,813 | 1.0 | <i>OLS</i> | 5,830,424 | 1.0 | <i>GAM</i> | 5,689,220 | 1.0 |
| | <i>SL</i> | 10,657,862 | 1.0 | <i>Lasso</i> | 5,830,718 | 1.0 | <i>OLS</i> | 5,734,502 | 1.0 |
| | <i>OLS</i> | 10,657,875 | 1.0 | <i>ElNet</i> | 5,830,867 | 1.0 | <i>Lasso</i> | 5,735,406 | 1.0 |
| | <i>GAM</i> | 10,657,875 | 1.0 | <i>RF</i> | 5,842,105 | 1.0 | <i>ElNet</i> | 5,735,468 | 1.0 |
| | <i>GLM l-n</i> | 10,657,875 | 1.0 | <i>Ridge</i> | 5,863,916 | 1.0 | <i>Ridge</i> | 5,766,763 | 1.0 |
| | <i>GLM l-g</i> | 10,657,875 | 1.0 | <i>GLM l-n</i> | 13,446,690 | 0.4 | <i>GLM l-n</i> | 41,287,682 | 0.1 |
| | <i>Ridge</i> | 10,658,065 | 1.0 | <i>GLM l-g</i> | Inc. value | - | <i>GLM l-g</i> | Inc. value | - |
| | Models ranking | MSE | RE | Models ranking | MSE | RE | Models ranking | MSE | RE |
| Fine granularity | <i>SL</i> | 10,618,384 | 1.0 | <i>SL</i> | 6,272,971 | 1.0 | <i>SL</i> | 6,100,482 | 1.0 |
| | <i>GLM l-n</i> | 10,620,560 | 1.0 | <i>GLM l-n</i> | 6,324,134 | 1.0 | <i>Lasso</i> | 6,130,363 | 1.0 |
| | <i>RF</i> | 10,629,964 | 1.0 | <i>Lasso</i> | 6,328,133 | 1.0 | <i>ElNet</i> | 6,130,485 | 1.0 |
| | <i>Lasso</i> | 10,632,277 | 1.0 | <i>ElNet</i> | 6,328,270 | 1.0 | <i>OLS</i> | 6,130,908 | 1.0 |
| | <i>ElNet</i> | 10,632,394 | 1.0 | <i>OLS</i> | 6,328,300 | 1.0 | <i>GAM</i> | 6,130,908 | 1.0 |
| | <i>OLS</i> | 10,632,491 | 1.0 | <i>GAM</i> | 6,328,300 | 1.0 | <i>Ridge</i> | 6,141,734 | 1.0 |
| | <i>GAM</i> | 10,632,491 | 1.0 | <i>Ridge</i> | 6,344,566 | 1.0 | <i>GLM l-n</i> | 6,164,559 | 1.0 |
| | <i>Ridge</i> | 10,632,648 | 1.0 | <i>RF</i> | 6,371,412 | 1.0 | <i>RF</i> | 6,175,096 | 1.0 |
| | <i>GLM l-g</i> | 10,633,779 | 1.0 | <i>GLM l-g</i> | 10,397,454 | 0.6 | <i>GLM l-g</i> | Inc. value | - |

Abbreviations: GAM, generalized additive model; MSE, mean squared error; OLS, Ordinary least squares; RF, random forest; SL, super learner.

7 | DISCUSSION

The recent developments in ML techniques to estimate healthcare spending have shown promising results, although as implemented in this study their performance is not uniformly dominant compared to OLS. The goal of this study is to provide a framework and toolkit for policymakers, therefore we selected well-established ML algorithms that required minimal tuning or could be implemented “off the shelf.” ML methods are still relatively unfamiliar for many policymakers, and some algorithms, like those based on neural networks, suffer from problems of interpretability that may limit their application (Kan et al., 2019). This study provided an in-depth analysis of the advantages and disadvantages of implementing ML techniques compared to standard regression (OLS) considering various scenarios of data granularity and range of variables and making use of healthcare administrative dataset from a large, unselected Italian population.

Our findings show that ML techniques, particularly SL, outperformed OLS in all data scenarios, although the adjusted-R² RE ranges between 0.02% points in the coarse granularity/poor range of variables scenario and 2.8% points in the coarse granularity/rich range of variables scenarios, indicating no statistically significant gain in our sample of 100,000 observations. Performance based on MSE also showed consistent results.

Concerning under/overcompensation measures, OLS showed the least, albeit negligible, overcompensation compared to ML techniques in the lowest observed spending percentiles in the coarse granularity/poor and fair range of variables setting. The top 1% spenders are better predicted with GAM or SL, except in the coarse granularity/poor range of variables and fine granularity/rich range of variables settings, where OLS was superior. The mean under/overcompensation also highlighted how the increase in the range of variables can considerably reduce the overestimation of the group of individuals in the 20th percentile and at the same time reduce the gap of the underestimation for the top of the spending distribution.

However, the gap between the best ML and OLS is relatively narrow, indicating that the implementation of more complex and sophisticated models does not lead to a significant increase in under/overcompensation reduction.

Unlike prior, US-based studies (McGuire et al., 2020; Rose, 2016), we did not find a consistent relationship between the increase in data granularity and/or range of variables and the advantage of ML over OLS. However, our most detailed data scenario (fine granularity, rich range of variables) included only 9 diagnosis-based risk adjustors, suggesting the ability of ML

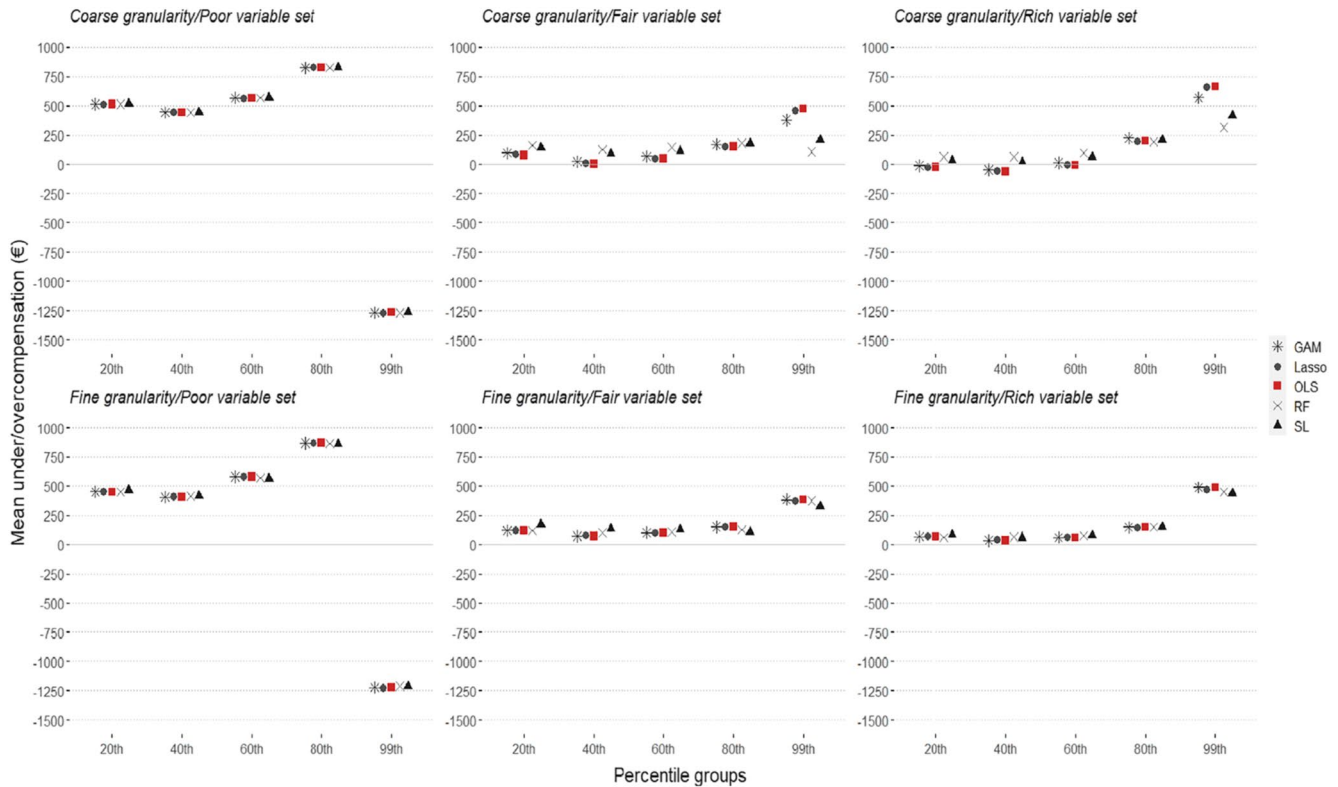


FIGURE 1 Under/overcompensation by quintiles for each data scenario. The ridge and elastic net mean compensations are not shown because they were very similar to lasso. GLMs are not shown because some of their mean compensation's values were completely out of range. Top 1% spenders were not included in the graph representation due to the higher range of values compared to quintiles (mean under compensation range from $-24,971$ to $-14,606$ €) [Colour figure can be viewed at wileyonlinelibrary.com]

algorithms to outperform OLS with fewer variables may be limited to settings with even more granular diagnosis groupings, such as the US formulas that use 90+ diagnosis-based risk adjustors.

Increasing the range of variables produced a significant improvement in the ability of all models (excepting the GLMs) to predict healthcare expenditure, more so in the coarse than fine granularity scenarios. The jump in the adjusted- R^2 from the coarse granularity/poor range of variables to the coarse granularity/fair range of variables due the addition of the number of hospitalizations in the risk-adjustment formula is greater than the increase moving to the fine granularity/rich range of variables scenario where PCG and DCG are added. We expected this result because the number of hospitalizations is much more predictive than demographic variables and it is highly correlated with the total expenditure in the concurrent approach. However, we encourage caution regarding its use in risk-adjustment models since it may lead to inappropriate treatment decisions (Geruso & Layton, 2015). Furthermore, the underfunding of high-cost morbidities that do not require hospitalization (e.g., outpatient dialysis for chronic kidney disease patients) may still present a challenge for policymakers.

Similar to McGuire, Zink, and Rose (McGuire et al., 2020), we find the addition of pharmaceutical-based indicators (moving from fair to rich range of variables in either granularity level) does not have a large impact on the adjusted- R^2 . Although the fine granularity variables DCG and PCG reduce the adjusted- R^2 , they may be regarded as reasonable trade-offs since they generally increase efficiency and fairness, and also reduce incentives for risk selection by including more risk-classes and are better predictors in prospective approaches.

Nevertheless, in addition to the upcoding behavior that morbidity-based indicators may induce, Geruso and McGuire (Geruso & McGuire, 2016) observed that HCC-based concurrent risk-adjustment systems increase fit but at the same time they reduce the power (the term power describes the share of costs at the margin born by the health plan (Geruso & McGuire, 2016) in cost containment by about 30%, meaning that concurrent diagnosis-based risk-adjusters weaken incentives for cost control. The inclusion of pharmaceutical-based indicators in the risk-adjustment model may incentivize health care providers to prescribe unnecessary medication (van Kleef et al., 2014) but, in the long-term, PCGs may direct providers' effort to cost-containing innovations (Beck et al., 2010).

We note some specific limitations to our study. There were many alternative fine granularities diagnosis risk-adjustors that we could have selected for this comparison. We limited our choice to DCG and PCG because they are well known in the risk-adjustment literature and commonly adopted for utilization and cost outcomes. However, we believe the methods we present can guide future comparisons of other risk-adjustors. Similarly, many other ML algorithms could be compared, for example, neural networks, other decision tree-based methods, or different choices could have been made regarding the tuning parameters (Hastie et al., 2009; James et al., 2013). Ultimately, our choice of algorithms was driven by a desire for both interpretability and potential familiarity for practitioners. Concerning data sources, the OPD includes only drug dispensations that are reimbursed by the NHS. The same caveat holds for OSD. Thus, we did not capture healthcare use paid by patients either directly or funded by private insurers. Finally, we relied on combining multiple data sources to select our base population, because we lacked access to the central Resident's Registry.

The choice of the model depends crucially on the performance measures; thus, it is of foremost importance to compare various measures of fit and interpret the results very carefully. It is more appropriate to rely on the adjusted- R^2 and MSE to maximize the variance explained and reduce the overall prediction error. However, if the primary goal is to minimize the differences between the predicted values and the observed values in specific subgroups that may be at risk of selection and to incentivize cost control, then the under/overcompensation or the PR are more targeted measures.

Policymakers should consider other evaluation criteria in designing the risk-adjustment model in addition to models' predictive performance, such as time/computing power, availability of data or appropriateness for incentives for risk selection and efficiency (van Veen et al., 2015).

Generally, ML techniques – particularly the SL – out-performed OLS. From an operational perspective, we suggest implementing ML methods to increase variance explained by the model in predicting the expenditure to further decrease risk selection, to reduce upcoding actions by selecting a subset of variables over all possible covariates, and to explore possible interaction terms with automatic processes (Rose, 2016). Retaining only a small number of important variables could reduce the opportunities to inflate the number of diagnosis and facilitate care management by emphasizing key risk-factors (Kan et al., 2019; Kronick & Welch, 2014). Conversely, if algorithm performance differences on the desired evaluation metrics are considered negligible and the investment in these tools too onerous in terms of computational time, OLS may be used with large populations. In our further analyses, when the entire population is considered rather than a subsample to estimate the OLS model, we indeed observed a narrowing of the gap in the MSE measure between ML models and the OLS.

To summarize, it is well known that in designing a risk-adjustment model it is important to evaluate not only the predictive performance but also the trade-offs with adverse economic behaviors (e.g., risk selection, moral hazard, cost containment). Our analysis shows a similar trade-off when deciding between the investment in increasing the range of variables and the granularity of the data, and in the search of the more appropriate statistical techniques to implement in conjunction with a given data scenario.

ACKNOWLEDGEMENT

The authors are grateful to the AUSL of Bologna for kindly providing the data and the financial support. The views expressed remain exclusively those of the authors. The usual disclaimers apply. Marica Iommi's grant was financed by the Local Health Authority (AUSL) of Bologna as part of the project “New funding mechanisms for the Italian National Health Service and the Emilia-Romagna Regional Health Service” jointly developed by the AUSL of Bologna and the Advanced School for Health Policies of the University of Bologna.

Open Access Funding provided by Università degli Studi di Bologna within the CRUI-CARE Agreement.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets generated and/or analyzed during the current study are property of a third party that is Emilia-Romagna Regional Health Agency (<https://assr.regione.emilia-romagna.it/>) and, although they are anonymized, datasets are not publicly available due to the current regulation on privacy. The description of the administrative databases is available from the website <https://salute.regione.emilia-romagna.it/siseps/sanita/asa/documentazione>. Other researchers can obtain access to the data through a formal request based on a research project to the Emilia-Romagna Regional Health Agency. We obtained the access to the data in the framework of a research agreement between the University of Bologna and the Local Health Authority of Bologna entitled “New funding mechanisms for the Italian National Health Service and the Emilia-Romagna Regional Health Service”.

ORCID

Marica Iommi  <https://orcid.org/0000-0003-4589-3474>

REFERENCES

- Armstrong, J. (2018). Health plan payment in Ireland. *Risk adjustment, risk sharing and premium regulation in health insurance markets* (pp. 331–364). Academic Press. <https://doi.org/10.1016/B978-0-12-811325-7.00012-9>
- Ash, A., Porell, F., Gruenberg, L., Sawitz, E., & Beiser, A. (1989). Adjusting Medicare capitation payments using prior hospitalization data. *Health Care Financing Review*, 10(4), 17–29. <http://www.ncbi.nlm.nih.gov/pubmed/10313277>
- Atella, V., Botti, R., Kopinska, J., & Marinacci, C. (2018). *L'allocation delle risorse in sanità: la situazione in Italia* (p. 24). Fondazione Farmafactoring.
- Bauhoff, S., Fischer, L., Göppfarth, D., & Wuppermann, A. (2017). Plan responses to diagnosis-based payment: Evidence from Germany's morbidity-based risk adjustment. *Journal of Health Economics*, 56, 397–413. <https://doi.org/10.1016/j.jhealeco.2017.03.001>
- Beck, K., Trottmann, M., & Zweifel, P. (2010). Risk adjustment in health insurance and its longterm effectiveness. *Journal of Health Economics*, 29(4), 489–498. <https://doi.org/10.1016/j.jhealeco.2010.03.009>
- Buchner, F., Wasem, J., & Schillo, S. (2017). Regression trees identify relevant interactions: Can this improve the predictive performance of risk adjustment? *Health Economics*, 26(1), 74–85. <https://doi.org/10.1002/hec.3277>
- Chini, F., Pezzotti, P., Orzella, L., Borgia, P., & Guasticchi, G. (2011). Can we use the pharmacy data to estimate the prevalence of chronic conditions? A comparison of multiple data sources. *BMC Public Health*, 11(1), 688. <https://doi.org/10.1186/1471-2458-11-688>
- Cid, C., Ellis, R., Vargas, V., Wasem, J., & Prieto, L. (2016). *Global risk-adjusted payment models*. https://doi.org/10.1142/9789813140493_0006
- Corrao, G., Rea, F., Di Martino, M., De Palma, R., Scondotto, S., Fusco, D., Lallo, A., Belotti, L. M. B., Ferrante, M., Pollina Addario, S., Merlino, L., Mancia, G., & Carle, F. (2017). Developing and validating a novel multisource comorbidity score from administrative data: A large population-based cohort study from Italy. *BMJ Open*, 7(12), e019503. <https://doi.org/10.1136/bmjopen-2017-019503>
- Dixon, J., Smith, P., Gravelle, H., Martin, S., Bardsley, M., Rice, N., Georghiou, T., Dusheiko, M., Billings, J., Lorenzo, M. D., & Sanderson, C. (2011). A person based formula for allocating commissioning funds to general practices in England: Development of a statistical model. *BMJ*, 343, d6608. <https://doi.org/10.1136/bmj.d6608>
- Ellis, R. P. (2007). *Risk adjustment in health care markets: Concepts and applications*. <https://doi.org/10.1002/9783527611294.ch8>
- Ellis, R. P., & Ash, A. (1995). Refinements to the diagnostic cost group (DCG) model. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, 32(4), 418–429. <http://www.ncbi.nlm.nih.gov/pubmed/8567079>
- Ellis, R. P., Martins, B., & Rose, S. (2018). Risk adjustment for health plan payment. In *Risk adjustment, risk sharing and premium regulation in health insurance markets* (pp. 55–104). <https://doi.org/10.1016/B978-0-12-811325-7.00003-8>
- Ellis, R. P., Pope, G. C., Iezzoni, L., Ayanian, J. Z., Bates, D. W., Burstin, H., & Ash, A. S. (1996). Diagnosis-based risk adjustment for medicare capitation payments. *Health Care Financing Review*, 17(3), 101–128. <http://www.ncbi.nlm.nih.gov/pubmed/10172666>
- Emilia-Romagna, R. (2014). *Tariffe DRG - allegato 3 DGR 1673/2014*.
- Eurostat. (2014). *Eurostat - degree of urbanisation (DEGURBA)*. Accessed 10 January 2021. <https://ec.europa.eu/eurostat/web/degree-of-urbanisation/data/database>
- Ferre, F., de Belvis, A. G., Valerio, L., Longhi, S., Lazzari, A., Fattore, G., & Maresso, A. (2014). Italy: Health system review. *Health Systems in Transition*, 16(4), 1–168. <http://www.ncbi.nlm.nih.gov/pubmed/25471543>
- Geruso, M., & Layton, T. (2015). *Upcoding: Evidence from medicare on squishy risk adjustment*. <https://doi.org/10.3386/w21222>
- Geruso, M., & McGuire, T. G. (2016). Tradeoffs in the design of health plan payment systems: Fit, power and balance. *Journal of Health Economics*, 47, 1–19. <https://doi.org/10.1016/j.jhealeco.2016.01.007>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. <https://doi.org/10.1007/978-0-387-84858-7>
- Henríquez, J., Iommi, M., McGuire, T. G., Mentzakis, E., & Paolucci, F. (2020). Designing feasible and effective health plan payments in countries with data availability constraints. *Journal of Health Economics*. (Manuscript submitted).
- Iezzoni, L. (2012). *Risk adjustment for measuring healthcare outcomes* (4th ed.). Health Administration Press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. <https://doi.org/10.1007/978-1-4614-7138-7>
- Jones, A. M. (2010). *Models for health care*. (No. 10/01).
- Kan, H. J., Kharrazi, H., Chang, H. Y., Bodycombe, D., Lemke, K., & Weiner, J. P. (2019). Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PLoS ONE*, 14(3), e0213258. <https://doi.org/10.1371/journal.pone.0213258>
- Kautter, J., Ingber, M., Pope, G. C., & Freeman, S. (2012). Improvements in Medicare Part D risk adjustment. *Medical Care*, 50(12), 1102–1108. <https://doi.org/10.1097/MLR.0b013e318269eb20>
- Kronick, R., & Welch, P. (2014). Measuring coding intensity in the Medicare advantage program. *Medicare & Medicaid Research Review*, 4(2). <https://doi.org/10.5600/mmrr.004.02.a06>
- Lamers, L. M. (1999). Pharmacy costs groups. *Medical Care*, 37(8), 824–830. <https://doi.org/10.1097/00005650-199908000-00012>
- McGuire, T., Zink, A., & Rose, S. (2020). *Simplifying and improving the performance of risk adjustment systems*. <https://doi.org/10.3386/w26736>
- Paolucci, F., Sequeira, A. R., Fouda, A., & Matthews, A. (2018). Health plan payment in Australia. In *Risk adjustment, risk sharing and premium regulation in health insurance markets* (pp. 181–208). <https://doi.org/10.1016/B978-0-12-811325-7.00006-3>
- Pope, G. C., Ellis, R. P., Wu, B., & Ash, A. S. (2000). *Final report: Diagnostic cost group hierarchical condition category models for medicare risk adjustment*.

- Radinmanesh, M., Ebadifard Azar, F., Aghaei Hashjin, A., Najafi, B., & Majdzadeh, R. (2021). A review of appropriate indicators for need-based financial resource allocation in health systems. *BMC Health Service Research*, 21(1), 674. <https://doi.org/10.1186/s12913-021-06522-0>
- Rose, S. (2016). A machine learning framework for plan payment risk adjustment. *Health Services Research*, 51(6), 2358–2374. <https://doi.org/10.1111/1475-6773.12464>
- Schmid, C. P. R., & Beck, K. (2016). Re-insurance in the Swiss health insurance market: Fit, power, and balance. *Health Policy*, 120(7), 848–855. <https://doi.org/10.1016/j.healthpol.2016.04.016>
- Skrami, E., Carle, F., Villani, S., Borrelli, P., Zambon, A., Corrao, G., Trerotoli, P., Guardabasso, V., & Gesuita, R. (2019). Availability of real-world data in Italy: A tool to navigate regional healthcare utilization databases. *International Journal of Environmental Research and Public Health*, 17(1), 8. <https://doi.org/10.3390/ijerph17010008>
- Smith, P. (2006). *Formula funding of public services*. Routledge.
- Trottmann, M., Telsler, H., Stämpfli, D., Hersberger, K., Matter, K., & Schwenkglens, M. (2015). *Übertragung der niederländischen PCG auf schweizer verhältnisse: Schlussbericht*. Bern.
- Trottmann, M., Weidacher, A., & Leonhardt, R. (2010). *Morbiditätsbezogene Ausgleichsfaktoren im Schweizer Risikoausgleich: Gutachten im Auftrag des Bundesamts für Gesundheit*. Verisk Health.
- van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). *Super Learner Statistical applications in genetics and molecular biology*, 6(1). <https://doi.org/10.2202/1544-6115.1309>
- van der Laan, M. J., & Rose, S. (2011). *Targeted learning*. <https://doi.org/10.1007/978-1-4419-9782-1>
- van Kleef, R. C., Eijkenaar, F., van Vliet, R. C. J. A., & van de Ven, W. P. M. M. (2018). Health plan payment in The Netherlands. In *Risk adjustment, risk sharing and premium regulation in health insurance markets* (pp. 397–429). <https://doi.org/10.1016/B978-0-12-811325-7.00014-2>
- van Kleef, R. C., van Vliet, R. C. J. A., & van Rooijen, E. M. (2014). Diagnoses-based cost groups in the Dutch risk-equalization model: The effects of including outpatient diagnoses. *Health Policy*, 115(1), 52–59. <https://doi.org/10.1016/j.healthpol.2013.07.005>
- van Veen, S. H. C. M., van Kleef, R. C., van de Ven, W. P. M. M., & van Vliet, R. C. J. A. (2015). Improving the prediction model used in risk equalization: Cost and diagnostic information from multiple prior years. *The European Journal of Health Economics*, 16(2), 201–218. <https://doi.org/10.1007/s10198-014-0567-7>
- van Veen, S. H. C. M., van Kleef, R. C., van de Ven, W. P. M. M., & van Vliet, R. C. J. A. (2018). Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees. *Health Economics*, 27(2), e1–e12. <https://doi.org/10.1002/hec.3523>
- Velasco, C., Henríquez, J., & Paolucci, F. (2018). Health plan payment in Chile. In *Risk adjustment, risk sharing and premium regulation in health insurance markets* (pp. 235–261). <https://doi.org/10.1016/B978-0-12-811325-7.00008-7>
- Wagner, T. H., Upadhyay, A., Cowgill, E., Stefos, T., Moran, E., Asch, S. M., & Almenoff, P. (2016). Risk adjustment tools for learning health systems: A comparison of DxCG and CMS-HCC V21. *Health Services Research*, 51(5), 2002–2019. <https://doi.org/10.1111/1475-6773.12454>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Iommi, M., Bergquist, S., Fiorentini, G., & Paolucci, F. (2022). Comparing risk adjustment estimation methods under data availability constraints. *Health Economics*, 31(7), 1368–1380. <https://doi.org/10.1002/hec.4512>