ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Predicting Frailty Condition in Elderly Using Multidimensional Socioclinical Databases

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Bertini, F., Bergami, G., Montesi, D., Veronese, G., Marchesini, G., Pandolfi, P. (2018). Predicting Frailty Condition in Elderly Using Multidimensional Socioclinical Databases. PROCEEDINGS OF THE IEEE, 106, 1-15 [10.1109/JPROC.2018.2791463].

Availability:
This version is available at: https://hdl.handle.net/11585/619981 since: 2018-02-07

Published:

DOI: http://doi.org/10.1109/JPROC.2018.2791463

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

26 June 2024

This is the final peer-reviewed accepted manuscript of:

F. Bertini, G. Bergami, D. Montesi, G. Veronese, G. Marchesini and P. Pandolfi, "Predicting Frailty Condition in Elderly Using Multidimensional Socioclinical Databases," in *Proceedings of the IEEE*, vol. 106, no. 4, pp. 723-737, April 2018.

The final published version is available online at: http://dx.doi.org/10.1109/JPROC.2018.2791463

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Predicting Frailty Condition in Elderly Using Multi-Dimensional Socio-Clinical Databases

Flavio Bertini, Giacomo Bergami, Danilo Montesi, Giacomo Veronese, Giulio Marchesini and Paolo Pandolfi

***Abstract*—Smart Cities face the challenge of combining sustainable national welfare with high living standards. In the last decades life expectancy increased globally, leading to various age-related issues in almost all developed countries. Frailty affects elderly who are experiencing daily life limitations due to cognitive and functional impairments and represents a remarkable burden for national health systems. In this paper we proposed two different predictive models for frailty by exploiting 12 socio-clinical databases. Emergency hospitalization or all-cause mortality within a year were used as surrogates of frailty. The first model was able to assign a *frailty risk* score to each subject older than 65 years old, identifying 5 different classes for tailor made interventions. The second prediction model assigned a *worsening risk* score to each subject in the first non-frail class, namely the probability to move in a higher frailty class within the year. We conducted a retrospective cohort study based on the whole elderly population of the Municipality of Bologna, Italy. We created a baseline cohort of 95,368 subjects for the *frailty risk* model and a baseline cohort of 58,789 subjects for the *worsening risk* model, respectively. To evaluate the predictive ability of our models through calibration and discrimination estimates, we used respectively a six-year and a four-year observation period. Good discriminatory power and calibration were obtained, demonstrating a good predictive ability of the models.**

***Index Terms*—Smart Healthcare, Smart City, Healthcare Data Analysis, Frailty Condition, Aging Society, Predictive Models.**

I. Introduction

Smart City is a broad concept mainly encompassing the process through which cities become more liveable by combining different smart layers, like people, mobility, governance, economy, environment and living [1], [2]. The challenge of combining sustainable national welfare with high living standard is pivotal in the context of an ageing society. The ageing phenomenon began fifty years ago due to a combination of a higher life expectancy and a lower birth rate (Figure 1). The trend is now quickly accelerating worldwide and according to 2015 United Nations reports [3], the number of people older than 65 years old is projected to reach 16.5% of the total world population by 2050. The elderly population is becoming a meaningful challenge for every national welfare system, in terms of services and costs [4], [5], as the health spending increases considerably in higher age classes. In 2011, according to Medicare program data [6], beneficiaries older than 65 years old requested 78% of the total Medicare spending (Figure 2).

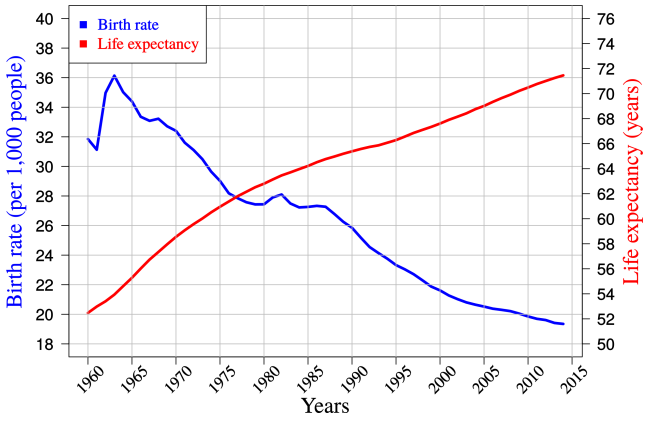

Fig. 1. Birth ratio versus life expectancy, world population trends from 1960 (source: World Bank Open Data).

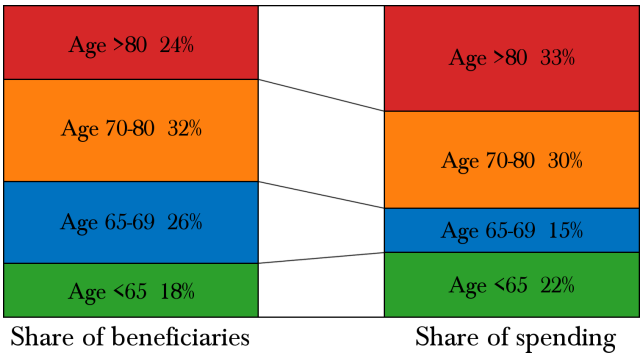

Fig. 2. Medicare beneficiaries versus Medicare spending in 2011 (source: Medicare, the United States national social insurance program).

In this scenario, Smart City has to face elderly people needs at different levels, in order to make the whole environment more comfortable for the ageing population [7], [8]. In particular, a Smart City can play a pivotal role in preventing adverse outcomes in elderly people [9], by adequately planning health and social care through the adoption of early screening and

F. Bertini, "L. Galvani" Interdepartmental Centre, University of Bologna and Department of Computer Science and Engineering, University of Bologna, Bologna, 40126 Italy (e-mail: flavio.bertini2@unibo.it).

G. Bergami and D. Montesi, Department of Computer Science and Engineering, University of Bologna, Bologna, 40126 Italy (e-mail: giacomo.bergami2@unibo.it; danilo.montesi@unibo.it).

G. Veronese, Department of Emergency Medicine, ASST Grande Ospedale Metropolitano Niguarda, University of Milano-Bicocca, Milan, Italy and Unit of Metabolic Diseases & Clinical Dietetics, Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy (e-mail: veronese.giacomo@gmail.com).

G. Marchesini, Unit of Metabolic Diseases & Clinical Dietetics, Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy (e-mail: giulio.marchesini@unibo.it).

P. Pandolfi, Epidemiology and Health Promotion Unit, Department of Public Health, AUSL Bologna, Bologna, Italy (e-mail: paolo.pandolfi@ausl.bologna.it).

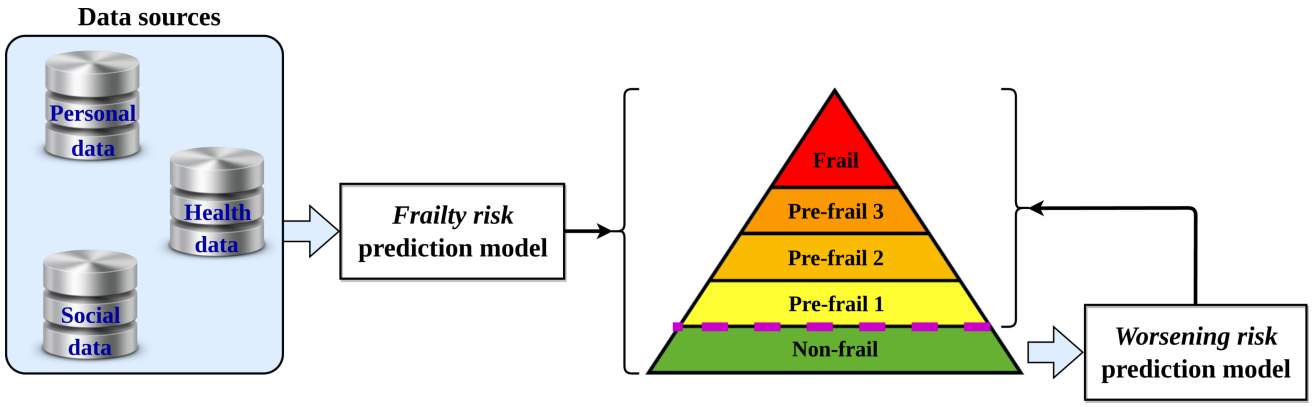

Fig. 3. Functional scheme of the *frailty risk* and *worsening risk* models. The *frailty risk* model stratifies subjects older than 65 years old in five classes, while the *worsening risk* model further classifies non-frail subjects (identified by the purple dotted line on the bottom of the triangle) according to their risk of becoming frail.

constant monitoring approaches [10]. The value of these strategies is twofold: mitigate the increasing health costs for elderly people and prevent negative events that might compromise their autonomy. For instance, the assessment of the elderly health condition allows to promote primary, secondary and tertiary prevention strategies[1], such as countermeasures under adverse weather conditions and customized services including personal transportation, health counseling, pharmaceutical assistance program and family caregiver support programs.

Frailty is one of the most crucial and emerging age-related conditions that generally represents an increasing limitation in daily activities. Typically, frailty results from functional decline and cognitive impairment, both contributing to develop a higher sensitivity to minor stressor events. Usually, this condition tends to get gradually worse over time, leading to several adverse outcomes encompassing disability, institutionalization, hospitalization and death [11], [12], [13]. Recent evidence supports the concept of frailty defined as the cumulative effect of heterogeneous deficits based on several socio-clinical variables [14], [15], [16].

Since frailty may be a reversible condition [17], early screening is of utmost importance in order to deliver preventive and tailored interventions [18]. In this paper we proposed two frailty predictive models for subjects older than 65 years old by exploiting information from 12 socio-clinical databases available in the Municipality of Bologna (Figure 3). We firstly created a data warehouse, namely a multi-dimensional database, combining 12 different socio-clinical data sources with information form 2009 to 2016. The data warehouse allowed to clean and integrate the data to be used in the proposed models. The first model characterized elderly people according to expected *frailty risk*[2], namely the probability of hospitalization or death within a year, extending the frailty definition proposed in [19]. The *frailty risk* model encompassed 27 clinical and socio-economic variables and was used to stratify subjects older than 65 years old into five risk classes, where the first class identified non-frail subjects. The risk stratification in five classes was proposed to help differentiating healthcare interventions. The lower the assigned *frailty risk* score, the lesser the probability that hospitalisation or death occurred within a year. The second model assigned a *worsening risk* to each subject in the lowest class, namely "non-frail" class. The *worsening risk* represented the probability to become frail, that was the risk for non-frail subjects to move in a higher frailty class within the year. This model encompassed 26 clinical and socio-economic variables. The retrospective cohort studies were conducted on the whole elderly population of the Municipality of Bologna (380,181 residents in 2010[3]). The over 65 years old category (25.93% of the overall population) represented the baseline cohort. In particular, we created a baseline cohort of 95,368 subjects from 2009 to 2010 for the first model and a new independent baseline cohort of 58,789 subjects from 2011 to 2012 for the second model. Then, we used a six-year (2011-2016) and a four-year (2013-2016) observation period[4] to assess the predictive ability of the *frailty risk* and *worsening risk* model, respectively.

The strengths of our study include its population-based design, the use of high quality routinely collected data from socio-clinical databases and the inclusion of several socio-clinical variables, that is a step forward in the field of frailty predictive modelling. Moreover, both validated models allow to categorize frailty and assess its severity.

The present work is organized as follows. In Section II, we reviewed the literature on three different topics: frailty condition in elderly people, data warehouse and predictive models for health. In Section III, we described the creation process of the data warehouse. In Section IV, we provided a brief background in logistic regression to help understanding the prediction models construction process. The baseline cohort and the results for the *frailty risk* model were presented in Section V. A comparison among different prediction models for frailty was discussed in Section VI. We presented the *worsening risk* model, including the baseline cohort and the

[1]Primary, secondary and tertiary prevention aim to prevent diseases, reduce the impact of ongoing diseases and soften the impact of diseases that have lasting effects, respectively.

[2]A prototype that implements the *frailty risk* calculator can be tested here: http://smartdata.cs.unibo.it/frailtycalc/

[3]According to the Italian National Institute of Statistics.

[4]The whole dataset is available on demand from the corresponding author.

results, in Section VII. Concluding remarks were made in Section VIII.

II. Related Works

In this section, we discussed the available literature on three different topics, all relevant for a full comprehension of our work. Firstly, in Section II-A, we focused on frailty in elderly people. Next, in Section II-B, we described several approaches proposed for data warehousing within the health setting. In the end, in Section II-C, we discussed predictive models in healthcare.

A. Frailty in elderly people

Elderly people generally develop a wide variety of age-related conditions that contribute to increase their vulnerability to minor stressor events and lead to loss of autonomy. The phenomenon is well-known as frailty, nevertheless its concept has not yet emerged as a well-defined clinical or social concept. In the past, clinicians used to separate the concepts of frailty, comorbidity and disability [20] and, initially, they proposed frailty models exclusively based on biological and clinical factors [11], [21]. Recent reviews tend to identify frailty as a complex interplay of a wide variety of heterogeneous factors, including not only clinical aspects but also socio-economic conditions [19], [22], recently demonstrated to influence frailty progression [14], [15], [23]. In [24], the authors included also psychological, social and environmental factors to better define frailty. Additionally, the relation between cognitive decline and frailty condition was widely discussed in [25]. Socio-clinical and administrative databases represent a potentially ideal source to implement models able to detect frailty and measure its severity [12]. In [26], the authors proposed a logistic regression based model for cardiac surgery patients, in order to reduce the mortality ratio and prolonged institutional care risk. However, few models aimed at stratifying the elderly population according to the estimated frailty risk have been validated. Moreover, they were primarily derived from a single source of information, like primary health care data [27], pharmacoepidemiologic data [28] or in-hospital data [16]. According to the most recent literature [29], [30], we validated two models by exploiting a wide range of socio-clinical variables to detect and measure frailty.

B. Healthcare data warehouse

The healthcare setting is generally perceived as being "data rich", since its datasets are daily updated with clinical and administrative patient data. Data warehouse systems are needed for preprocessing operations (e.g., data integration and cleaning) and in order to create a consistent dataset ready to be mined. The aim is to extract useful information and improve healthcare quality [31]. The value of a data warehouse system in healthcare was initially examined in [32]. Moreover, in [33] the authors discussed the role of data warehousing in an academic medical center, including research and education purposes in addition to administration and management. In [34], data mining techniques were applied to a diabetic data warehouse in order to investigate clinical topics and improve administrative management in diabetic patients. A clinical data warehouse importing data from three public hospitals was described in [35]. The authors proposed models to study antimicrobial resistance and to monitor hospital-acquired bloodstream infections and costs. Recently, data warehousing combined with data mining techniques in health domain have been widely discussed. In [36], the authors examined the usefulness of a data warehouse system integrating laboratory, administrative and clinical data, in order to develop further data mining techniques. An interesting study investigating the benefit of a data warehousing technology in the radiation oncology field was proposed in [37]. In conclusion, data warehousing techniques have proven to be promising for healthcare information systems in several clinical fields, like intensive care [38] and clinical pathology [39]. We thus created a data warehouse that combines social, clinical and administrative data sources to assess and measure frailty in elderly people through prediction models.

C. Predictive models in healthcare

Predictive models and data mining are producing a revolution in healthcare, as described in [40]. Detecting adverse clinical events, reducing mortality rates and mitigating healthcare costs, all represent challenging tasks for researchers [41], [42]. In [43], the authors discussed data mining techniques in major healthcare areas such as evaluation of treatment effectiveness, healthcare management and fraud and abuse detection. The effectiveness of data mining in prediction and decision making in healthcare was widely examined in [44]. In [45], the authors proposed a hybrid approach by combining genetic algorithm and logistic regression techniques to predict the Alzheimer's disease progression. A model able to predict risk in developing a target disease was presented in [46], where the authors described a general top-k stability selection method in comparison to three classic classification methods (support vector machine, logistic regression and random forest), in order to select features from the electronic health records. In [47] and [48] different data mining and machine learning techniques were studied. In particular, the authors provided a detailed description of a decision tree and single and hybrid data mining techniques to characterize metabolic syndrome patients and diagnose heart disease. Interesting results were presented in [49], where a model based on rotation forest with alternating decision tree to assign a 5 year life expectancy index to subjects older than 50 years was built. Recently, social media provided another prosperous source of information that can be useful to extract data on public health monitoring. Some interesting attempts in this direction were presented in [50] and [51]. In the first work, the authors tried to predict the seasonal flu epidemics by building a model based on Google search queries. In the second one, the flu epidemic was monitored through Twitter using a support vector machine. Several methodologies can be used to build prediction models in healthcare and the choice significantly depends on the outcomes to be predicted. A comprehensive survey was done in [52] and [53]. In our work we used the logistic regression, as

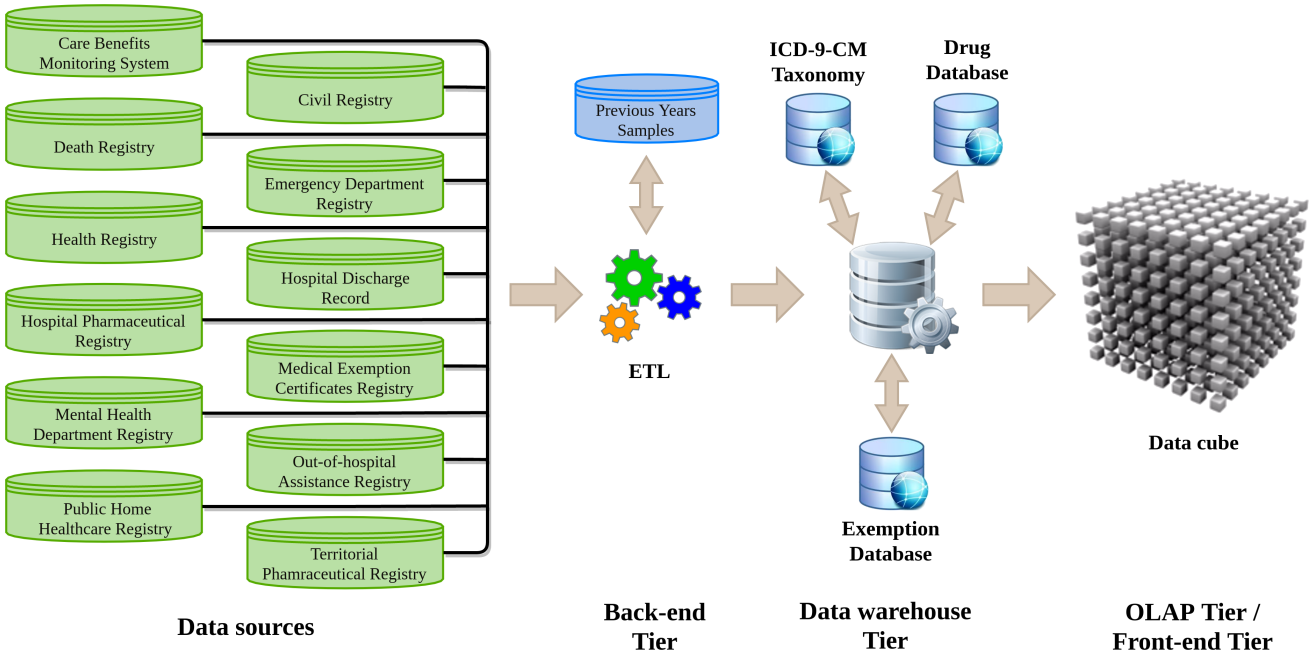

Fig. 4. The data warehouse architecture.

it is known to be less sensible to noisy data and as it allows to associate a continuous response (i.e., probability) to the dependent variable to be predicted.

III. Data Warehousing

In this section, we described the five-phase construction process of the data warehouse (Figure 4), allowing us to load, clean and integrate data from 12 socio-clinical databases [54].

Phase 1 - Data sources: The 12 data sources collect continuous, categorical and Boolean variables, including information on clinical and socio-economic aspects and health service resources utilization.

1) Care Benefits Monitoring System (CBMS) records information about public benefits.
2) Civil Registry (CR) collects socio-economic data characterizing each resident.
3) Death Registry (DR) tracks the deaths occurring in Bologna.
4) Emergency Department Registry (EDR) collects all admissions to emergency department.
5) Health Registry (HR) provides vital statistics and health data of the subjects.
6) Hospital Discharge Record (HDR) is built upon the hospitalization records and includes the *ICD-9-CM*[5] codes related to diagnosed diseases and medical interventions.
7) Hospital Pharmaceutical Registry (HPR) collects information and *Anatomical Therapeutic Chemical Classification System* (*ATC*) codes on all drugs directly prescribed by health care institutions.
8) Medical Exemption Certificates Registry (MECR) provides information and *Exemption Code* (*EC*) on all released medical exemption certificates.
9) Mental Health Department Registry (MHD) records subjects followed-up by the department of Mental Health.
10) Out-of-hospital Assistance Registry (OAR) contains all the health services provided to out-of-hospital patients.
11) Public Home Healthcare Registry (PHHR) provides information on home-based care services.
12) Territorial Pharmaceutical Registry (TPR) collects information and *ATC* codes on drug prescriptions presented at the dispensing pharmacy.

All the available raw data are listed in Table IV in the Appendix.

Phase 2 - Back-end Tier: In this phase the data retrieved from remote servers and off-line records was extracted, transformed and loaded together using solutions proposed in [55]. The ETL (Extract, Transform, Load) component produced an intermediate representation.
Firstly, the identification of each subject among the different data sources, namely data linking, was carried on by using the *Fiscal Code*[6]. Civil Registry and Health Registry represented the *Fiscal Code* ground truth. Then, the rest of the raw data was extracted from all the others data sources by using the *Fiscal Code*. Inconsistent values, like missing or invalid characters in date and place of birth, name and surname, were adjusted and replaced with valid ones by cross-checking all data sources.
In this phase, it was not possible to extract information nested in the text fields of the data sources. As a reason, in the Data warehouse Tier phase we integrated the available data

[5]International Classification of Diseases, 9^{th} revision, Clinical Modification. At the time of writing, Italian National Health System did not use the 10^{th} revision.

[6]The Fiscal Code is similar to the Social Security Number in the US or the National Insurance Number in the UK.

with other knowledge bases, like *ICD-9-CM Taxonomy*, *Drug Databases* and *Exemption Code Databases* in order to refine the dataset.

Phase 3 - Data warehouse Tier: The 12 data sources offered several variables, in some cases with missing values or redundancies. In this section, we discussed the operations through which the dataset was corrected, validated and enriched by using the following three main knowledge bases:

1) *ICD-9-CM Taxonomy* - The International Classification of Diseases (ICD) is an international standard classification system used to code diseases, symptoms, injuries and medical procedures. Assigning a standard code to each clinical case is intuitively a great advantage for statistical, epidemiological and policy-making purposes. In Italy, at the time of writing, the *ICD-9th revision-Clinical Modification* (*ICD-9-CM*) was still used. However, since *ICD-9-CM* does not provide a complete correlation between all the possible codes, a taxonomy was semi-automatically extracted from several clinical textbooks [56]. For each clinical variable previously selected, the taxonomy allows to reconcile the *ICD-9-CM* codes in our knowledge base with the codes used in the Hospital Discharge Record. The "Lin distance" [57], a well-known taxonomy distance measure of correctness, was used to select only the correct codes (see Table I).
2) *Drug Database* - The Anatomical Therapeutic Chemical (*ATC*) is a pharmaceutical coding system that allows to classify active ingredients of drugs. In order to map each *ATC* code into the right drug-related variable, we created a database matching the *ATC* codes and the related disease. Through this process, it was possible to enrich the previous *ICD-9-CM* information related to each selected variable (see Table I).
3) *Exemption Code Database* - In Italy, medical exemptions are coded by using both regional and national coding systems. The first one is a four alphanumeric code while the second is a three digits code. The Medical Exemption Certificates Registry adopts both coding systems. In order to match the right exemption code into the right disease, we created a database including both the regional and national coding system, further refining the information related to each selected variable (see Table I).

The three knowledge bases and the information redundancy among all data sources allowed to correct partial and incomplete data. In particular, for social variables we adopted the *mean* and *cold-deck* imputations through which each missing value was replaced with the mean of the observed values or by selecting the value from another source. For each clinical variable in Table I, we used imputation based on logical rules for the missing values. For instance, if a subject did not have any recent diagnosis of cancer, the respective variable could be set by using the *ICD-9-CM* codes for benign and malignant tumours, the *ATC* codes related to post-surgical removal treatments and tumours-related exemptions currently active for the subject. Since the *ICD-9-CM* taxonomy has a hierarchical structure and some categories include set of similar diseases, we grouped similar comorbidities. All the created groups are listed in Table V in the Appendix. In practice, we were able to intercept different disease forms using the *ICD-9-CM* codes, detect both initial and chronic stages of the pathology through the drugs knowledge base and identify the current health status using exemptions codes.

TABLE I
THE CODES FROM THREE DIFFERENT KNOWLEDGE BASES (*ICD-9-CM Taxonomy*, *Drug Database* AND *Exemption Code Database*) USED TO DETERMINE THE VALUE OF THE RELATIVE CLINICAL VARIABLES.

Variable	Databases	Codes extracted
Arthritis	MECR	*EC*: 006
Cancer	HDR HPR MECR	*ICD-9-CM*: 140.00-239.99 *ATC*: L01-L04 *EC*: 048
Cerebrovascular disease	HDR MECR	*ICD-9-CM*: 430.00-438.99 *EC*: 0B02
Deaf, mute and/or blind	MECR	*EC*: C05, C06
Dementia	HDR MECR	*ICD-9-CM*: 290 *EC*: 029, 011
Diabetes	HDR TPR HPR MECR	*ICD-9-CM*: 250 *ATC*: A10A, A10B, A10X *EC*: 013
Disability	MECR	*EC*: C01-C03, L01-L04, G01-G02, S01-S03, N01
Chronic kidney disease	HDR MECR	*ICD-9-CM*: 585.00-586.99 *EC*: 023
Gastric disease	MECR	*EC*: 009
Hyper-cholesterolemia	HDR MECR	*ICD-9-CM*: 270.0-272.2 *EC*: 025
Hypertension	HDR MECR	*ICD-9-CM*: 401.0-405.99 *EC*: 0A31, 0031
Liver disease	HDR MECR	*ICD-9-CM*: 571.2, 571.4-571.6, 571.8-571.9 *EC*: 008, 016
Parkinson's disease	HDR TPR HPR MECR	*ICD-9-CM*: 332 *ATC*: N04A, N04B *EC*: 038
Psychiatric disorders	TPR HPR MECR	*ATC*: N05, N06 *EC*: 044, 005
Respiratory diseases	HDR MECR	*ICD-9-CM*: 490.00-496.99 *EC*: 007, 024
Thyroid diseases	HDR TPR HPR	*ICD-9-CM*: 240.00-243.99, 245.00-246.99 *ATC*: H03 *ATC*: A10A, A10B, A10X
Vascular disease	HDR MECR	*ICD-9-CM*: 410.00-414.99, 428 *EC*: 0A02, 0C02, 021

Phase 4 - OLAP Tier: In data warehouse systems, OnLine Analytical Processing (OLAP) represents a specific approach in order to answer multi-dimensional analytical queries. OLAP is a well-known decision support tool and is characterized by aggregated and historical data, stored in multi-dimensional structure, namely *data cube*. Data mining techniques widely use the *data cube* to analyse multi-dimensional data from multiple data sources. In our case, a *data cube* was created according to the previously described phases. The three dimensions of our study were the subjects involved in the study, the time in years from 2009 to 2016 and all the variables collected from the selected data sources. However, since the techniques used in this work required a tabular representation of the *data cube*, we created a two-dimensional representation. In practical terms, for every year we provided a table where each row represented a subject and each column represented a variable.

Phase 5 - Front-end Tier: In this phase, the *data cube* was processed applying both statistical and data mining techniques. In our work, this phase coincided with the creation of the two predictive models (Sections V and VII).

IV. Logistic regression: background

In the current section, we provided a brief background in logistic regression in order to better understand the rationale for its use for the frailty prediction models.

Multivariate regression analysis detects the correlation between a set $X_1, \ldots, X_n$ of independent variables (i.e., the predictors) and a dependent variable E (i.e., the expected event) [58]. This technique has the following three characteristics: *i*) it can mix categorical and continuous predictors; *ii*) it is the most appropriate to solve binary classification tasks; *iii*) it is less likely to be influenced by noisy data [53].

Logistic regression is a specific type of multivariate regression where each row $X_1, \ldots, X_n$ is intercepted by a sigmoid function and the outcome is a probability function $P(E|X_1, \ldots, X_n)$. The training phase based on maximum likelihood technique assigns a regression coefficients β_i to each predictor $X_1, \ldots, X_n$. The $\beta_1, \ldots, \beta_n$ coefficients punctually describe the strength of each predictor. Moreover, logistic regression allows to estimate the influence of each single predictor in intercepting the expected event E. In particular, variables shown to be significantly associated with the outcome by univariate analysis are selected for multi-dimensional analysis.

Logistic regression is the simplest and most interpretable binary classifier. We mainly used it to distinguish non-frail and frail subjects and not at risk and at risk subjects in the first and second model, respectively. In the first *frailty risk* model the expected event E was defined as an emergency hospitalization or the all-cause mortality within a year. We calculated the probability of E to identify non-frail and frail subjects older than 65 years old. Moreover, since the model was thought to be used in order to enable preventive and tailored interventions in a real-life healthcare service, we also stratified the subjects into five risk classes according to the probability of E. In the second *worsening risk* model, the expected event E was defined as the probability to become frail within the year. In other words, E was the progression towards a higher frailty class within the year for the subjects belonging to the non-frail class. In this case, the probability of E was used to distinguish non-frail subjects identified by the first model into not at risk and at risk of evolving into frail classes.

Since the models worked in a data rich environment, both in terms of number of records and number of predictors, we decided to categorize continuous predictors (e.g., age, income) into classes. Thus, each new class was represented by a new *dummy variable* with an associated β_i. Even though not strictly necessary, this approach led to a better balance between accuracy and interpretability of the models, a relevant characteristic in socio-clinical settings.

V. Frailty risk model

In this section, we discussed how we developed the *frailty risk* prediction model through which every over 65 years old subject was assigned a frailty score according to the probability of emergency hospitalization or death within a year [19], [30]. As described above, logistic regression classified individuals into non-frail and frail subjects. Frail subjects were further stratified by their increasing risk of adverse outcomes.

A. Baseline cohort selection

First, we used the *data cube* to create a two-year anonymised cohort of subjects (from January 1^{st}, 2009 to December 31^{st}, 2010). A two-year observation period was selected in order to improve the detection of information through all the 12 data sources. All individuals younger than 63 years old on January 1^{st}, 2009, all dead subjects in the selected biennium, as well as those who emigrated or were non-residents or without health care coverage were removed from the dataset. This filtering process produced a baseline cohort of 95,368 subjects older than 65 years old on January 1^{st}, 2011. A total of 14,812 events were observed during the follow-up year. This cohort was used to build the *frailty risk* prediction model.

B. Model development

The selected cohort of 95,368 subjects was randomly split in a **training set** and a **test set**, which did not significantly differ in terms of characteristics. The training set of 63,579 subjects (2/3 of the total) was used to tune the parameters β_i. The test set of 31,789 subjects (1/3 of the total) was used to evaluate the prediction capabilities. Data from January 1^{st}, 2011 to December 31^{st}, 2011 (1-year follow up) was investigated for predictor variables associated with the risk of emergency hospitalization or death, the main outcome of interest. Using logistic regression, odds ratios (ORs, that is e^{β_i} for each variable) and 95% confidence intervals (CIs) were computed to evaluate the association between the selected variables and the outcome. Variables shown to be significantly associated with the outcome by univariate analysis (p-value[7] <0.05) were

[7]The p-value is used in the context of null hypothesis testing (i.e., reductio ad absurdum) and quantifies the statistical significance of evidence. The 0.05 is the standard cut-off.

selected for multivariate analysis.
More than fifty different variables for each subject (see Table IV) were extracted from the 12 data sources. Univariate analysis discarded the non-significant ones and identified 27 socio-clinical predictors listed in Table VI in the Appendix. The absence of condition was used as reference. All "Comorbidity" and "Benefits" variables were represented by Boolean values. Whereas, for each non-categorical variable we specified the *dummy variable* added. In particular, all continuous variables were characterized as follows:

- "Age" was grouped using 3-year age bands with the 65-67 class used as reference category;
- "Housing condition" included privately-owned (reference category), renting and other;
- "Level of education" included primary, secondary and higher (reference category);
- "Income" included >75k€/year (reference category), 36-75k€/year, 28-36k€/year, 15-28k€/year and <15k€/year, according to the Italian personal income tax classes;
- "Marital status" included married (reference category), single, widowed and divorced;
- "Deprivation index" included very rich (reference category), rich, medium, poor and very poor;
- "Hospitalizations" and "Emergency Room visit" included 0 (reference category), 1, 2 and >2 visits per year, respectively;
- "Emergency hospitalization" represented events occurring within 30 days before the follow-up period;
- "Diagnostic test" represented tests performed within 90 days before the follow-up period;
- "Polypharmacy" represented prescriptions received within 90 days before the follow-up period.

The *frailty risk* score was estimated from the linear prediction starting from the log odds of the final model, as:

$$frailty\_risk = 100 \cdot P(E|X_1, \ldots, X_{27}) \qquad (1)$$

The score was defined ranging from 0, null risk, to 100, full risk and each subject was assigned a value defining the probability to undergo emergency hospitalization or death within a year.

Finally, a total of 5 classes with increasing risk of adverse outcomes were defined as follows: "non-frail" class (0-14%), "pre-frail 1" class (15-29%), "pre-frail 2" class (30-49%), "pre-frail 3" class (50-79%), "frail" class (80-100%). The first class, identifying non-frail subjects, was defined using the intersection point between sensitivity and specificity curves (i.e., probability ≈ 0.140, $frailty\_risk$ = 14.0%), as shown in Figure 5 and in line with previous evidence [59]. Sensitivity measured the non-frail subjects who were correctly identified as not having the event, whereas specificity measured frail subjects who were correctly identified as having the event. Figure 5 shows both sensitivity and specificity curves according to different cut-off probability values. The remaining classes were defined according to epidemiologists' opinion and previous studies reported in the literature [22].

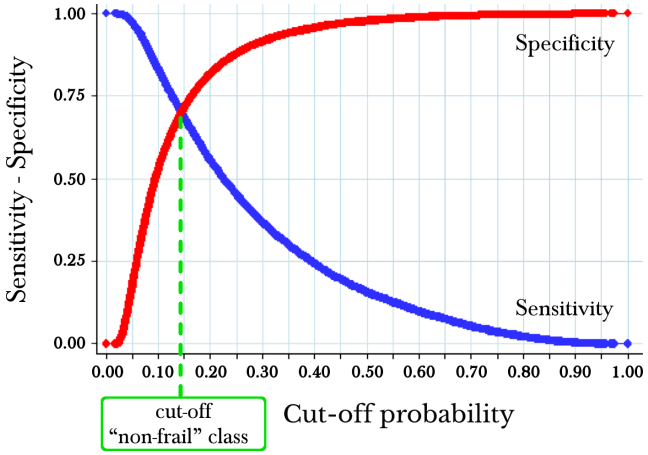

Fig. 5. Sensitivity and specificity curves related to the training set. The intersection point, that is a cut-off probability ≈ 0.140, is used as upper bound to define the "non-frail" class.

C. Results

The predictive ability of the *frailty risk* model was demonstrated through internal and external validations, using calibration and discrimination estimates.

The internal validation was performed by applying to the test set the regression coefficients $\beta_1, \ldots, \beta_{27}$ resulting from the training set. Then, for both training and test sets, the area under the receiver operating characteristic (AUROC) curve was computed to assess the discrimination capability of the model and the Hosmer-Lemeshow test was estimated to evaluate the goodness of fit. On the training set the AUROC value was 0.7681 and the Hosmer-Lemeshow 0.1099. On the test set these values were slightly lower: 0.6968 and 0.0768, respectively. Overall, the AUROC identified a good prediction and Hosmer-Lemeshow scores greater than 0.05 meant that differences among observed and expected events were not statistically significant, demonstrating a good discrimination capability and calibration of the *frailty risk* model.

The external validation was performed by cross matching the subjects in the "frail" class with those present in "Access Management of Integrated and Automated Social-Health Network" (GARCIA) [60], a local database composed of frail subjects identified by local health authorities through the "Brief Self-Sufficiency Index" (BINA) scale. The BINA is a multidimensional tool assessing the subject's functional and cognitive status, the availability of a caregiving network, as well as the housing and the neighbourhood quality [61]. The external validation showed that 100% of subjects in the GARCIA dataset belonged to the "frail" class according to our *frailty risk* prediction model.

Finally, even though there was no real data to validate the five classes, the appropriateness of the model was further assessed by illustrating the observed outcome rates occurring in a six-year follow-up period from 2011 to 2016 on the overall annual local population older than 65 years old. Table II shows the observed events in each class with the related percentage distributions and 95% CI computed using Poisson regression. The observed events are reported according to the expected risk class identified by the model. Since the number

TABLE II
OBSERVED EVENTS IN EACH CLASS WITH THE PERCENTAGE DISTRIBUTION AND 95% CI BETWEEN ROUND BRACKETS DURING A SIX-YEAR FOLLOW-UP PERIOD. THE OVERALL ANNUAL LOCAL POPULATION OLDER THAN 65 YEARS OLD IS SPECIFIED BETWEEN SQUARE BRACKETS UNDER EACH YEAR.

	Observed events in a six-year follow-up (% ± 95% CI)					
Risk classes	**2011** [104,128]	**2012** [99,455]	**2013** [99,823]	**2014** [99,920]	**2015** [99,831]	**2016** [100,314]
Non-frail (0-14)	5,256 (7.5 ± 0.2)	4,087 (6.7 ± 0.2)	4,356 (6.9 ± 0.2)	3,726 (6.2 ± 0.2)	3,600 (6.0 ± 0.2)	3,406 (5.7 ± 0.2)
Pre-frail 1 (15-29)	5,322 (23.4 ± 0.5)	5,046 (20.9 ± 0.5)	5,003 (21.4 ± 0.5)	4,653 (19.3 ± 0.5)	4,561 (18.9 ± 0.5)	4,428 (18.3 ± 0.5)
Pre-frail 2 (30-49)	2,961 (39.4 ± 1.1)	3,329 (36.7 ± 1.0)	3,261 (36.7 ± 1.0)	3,308 (34.4 ± 0.9)	3,460 (34.8 ± 0.9)	3,298 (31.8 ± 0.9)
Pre-frail 3 (50-79)	1,893 (57.1 ± 1.7)	2,395 (54.0 ± 1.5)	2,294 (53.2 ± 1.5)	2,557 (51.0 ± 1.4)	2,718 (51.6 ± 1.3)	2,606 (46.5 ± 1.3)
Frail (80-100)	241 (74.4 ± 4.7)	396 (77.0 ± 3.6)	292 (69.5 ± 4.4)	393 (69.0 ± 3.8)	462 (72.5 ± 3.5)	427 (59.0 ± 3.6)

of older than 65 years old subjects decrease during the six-year follow-up period, the "non-frail" and "pre-frail 1" classes show a decreasing trend. Whereas the remaining "pre-frail 2", "pre-frail 3" and "frail" classes show an increasing trend. In particular, from 2011 to 2015, for 4 out of 5 classes, the percentage of the observed events fell within the risk prediction range of the class. For instance, in 2011 among the subjects belonging to the first class, the percentage of the observed event was 7.5%, which falls within the range 0-14%. This holds true for the first three classes in 2016. The model slightly overestimated the events observed in the last "frail" class throughout the six-year follow-up, particularly in the long-term period. For instance, in 2011 among the subjects belonging to the "frail" class, the percentage of the observed events was 74.4%, lower than the minimum value of 80% defining the class itself. Similarly, the average percentage of the observed events during the follow up in the "frail" class was 70.2%. However, this overestimation can be considered safe and preferable, since allows healthcare services to early detect the most vulnerable subjects. In 2016 the events occurred in "pre-frail 3" and "frail" classes were noticeably overestimated, suggesting recalibration of the model due to social and economic changes occurring in a five-year period.

VI. COMPARISON OF FRAILTY PREDICTION MODELS

In this section we provided a comparison of different models that can be used for frailty prediction, in order to understand the advantages of logistic regression versus others available methods described in literature, including recursive partitioning (rpart) [62], generalized boosted [63], random forest [64], support vector machine (svm) both with polynomial and radial basis function (RBF) kernel [65]. All these methods allow to associate a continuous response to the expected event and a score can later be computed in order to stratify subjects in risk classes. The comparison was carried on testing the prediction capabilities of the different models. We used the previously described training and test dataset to discriminate non-frail and frail subjects.

In order to improve the performance, the parameters for each model were tuned using a grid search over supplied parameter ranges. The tuning phase selected the best parameters for each model minimizing the mean squared error through a 10-fold cross validation process. In particular, the tuning phase identified the following parameters for each model:

- *rpart* - the observations that had to exist in a node for a split to be attempted was 5;
- *generalized boosted* - the number of trees was 150, the maximum depth of variable interactions was 3, the shrinkage parameter applied to each tree in the expansion was 0.1 and the minimum number of observations in the trees terminal nodes was 10;
- *random forest* - the number of variables randomly sampled as candidates at each split was 3 and the number of trees to grow was 700;
- *svm polynomial* - the cost of constraints violation was 1 and the required parameter gamma was 0.05;
- *svm RBF* - the cost of constraints violation was 1 and the required parameter gamma was 0.1.

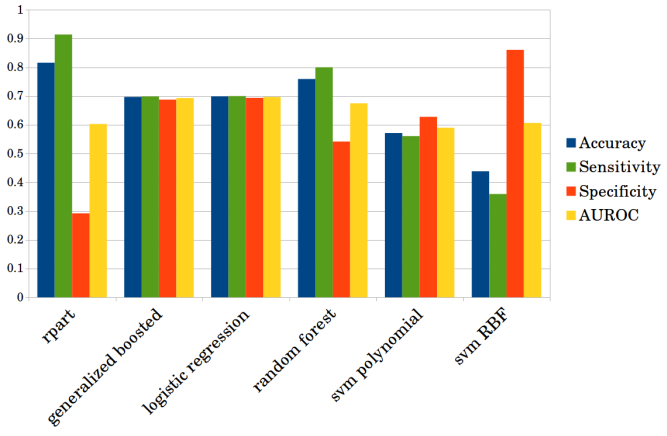

Fig. 6. Accuracy, sensitivity, specificity and AUROC results on the test set for the six compared models.

We used accuracy (i.e., the closeness of measurements to the true values), sensitivity, specificity and AUROC measures to compare the models. Figure 6 documents the results of the comparison. The worse accuracy result was obtained with support vector machine, both with polynomial and RBF kernel. Despite rpart and random forest presented higher values of accuracy in comparison to logistic regression, their low specificity results represented a high disadvantage. In other words, rpart and random forest failed to identify the most vulnerable subjects. Finally, logistic regression obtained slightly better results in comparison to generalized boosted.

VII. Worsening risk model

Results from the *frailty risk* model showed that elderly subjects, even the non-frail ones, typically move toward a frail class in the course of their life. Thus, early identification of future frail subjects is of utmost importance in order to adequately plan health and social care.

The second model proposed in this section attempted to identify non-frail individuals expected to become frail within a year. Therefore, we assigned a *worsening risk* score to each subject belonging to the "non-frail" class identified by the previous prediction model (Section V) to define his/her probability of becoming frail. This so called *worsening risk* model ideally allows appropriate planning of health resource for those predicted to become frail over the following year.

A. Baseline cohort selection

Since we needed to fit the *worsening risk* model for the non-frail population and the event E was the frailty class change within the year, we did not use the same training/test dataset of the previous model but we created a new two-year (from January 1^{st}, 2011 to December 31^{st}, 2012) anonymised cohort. By construction, the two cohorts did not share any records. All dead subjects in the biennium, as well as those who emigrated or were non-residents or without health care coverage were removed from the cohort. Moreover, using the outcome of the previous frailty prediction model, we only selected subjects in the "non-frail" class with an assigned *frailty risk* score. For this reason, the covering period of this baseline cohort was shifted forward of two years in comparison to the *frailty risk* prediction model and the training/test process was independent from the previous one. The filtering process produced a baseline cohort of 58,789 subjects older than 65 years old on January 1^{st}, 2013, where we observed 4,771 events during the follow-up year. This new baseline cohort was used to build the *worsening risk* prediction model.

B. Model development

The selected cohort of 58,789 subjects was randomly split in a **training set** and a **test set** which did not significantly differ in terms of characteristics. The training set of 39,193 subjects (2/3 of the total) was used to tune the parameters β_i. The test set of 19,596 subjects (1/3 of the total) was used to evaluate the prediction capability. Using data available on January 1^{st}, 2013, we searched for predictor variables associated with the risk to become frail. Using logistic regression, ORs and 95% CIs were computed to evaluate the association between the selected variables and the outcome. Variables shown to be significantly associated with the outcome (p-value <0.05) by univariate analysis were selected for multivariate analysis.
More than fifty different variables for each subject (see Table IV) were made available from the 12 data sources. Out of all the available variables, the univariate analysis discarded the non-significant ones and identified 26 socio-clinical predictors listed in Table VII in the Appendix. Compared with the variables set for the *frailty risk* prediction model, "Emergency hospitalization" was the only variable excluded.
The *worsening risk* score was again estimated from the linear prediction starting from the log odds of the final model, as:

$$worsening\_risk = 100 \cdot P(E|X_1, \ldots, X_{26}) \qquad (2)$$

However, in this case, only two classes were defined according to risk of becoming frail within the year: "non at risk" class (0-10.8%) and "at risk" class (10.8-100%).

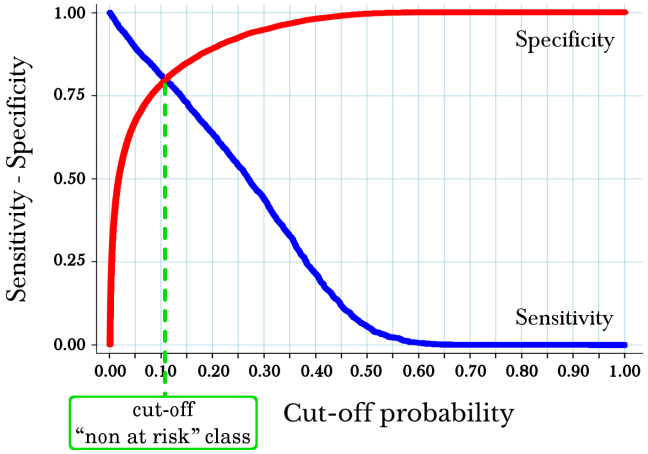

Fig. 7. Sensitivity and specificity curves related to the training set. The intersection point, that is a cut-off probability $\approx$ 0.108, is used as upper bound to define non-frail individuals non at risk of becoming frail.

Figure 7 shows the selection of the cut-off used to discriminate non at risk and at risk subjects. The "non at risk" class was defined using the intersection point between the sensitivity and specificity curves (i.e., probability $\approx$ 0.108, $worsening\_risk$ = 10.8%), in line with previous evidence [59].

C. Results

The predictive ability of the *worsening risk* model was demonstrated only through internal validation, since no external dataset was available for external validation. In particular, for each year we verified the number of non-frail subjects becoming frail within a year according to the initially described *frailty risk* prediction model.

The internal validation was performed by applying to the test set the regression coefficients $\beta_1, \ldots, \beta_{26}$ resulting from the training set. Then, for both training and test sets, the AUROC curve was computed to assess the discrimination capability of the model. Moreover, the Brier score test was

TABLE III
OBSERVED EVENTS IN BOTH CLASSES WITH THE PERCENTAGE DISTRIBUTION AND 95% CI BETWEEN ROUND BRACKETS DURING A FOUR-YEAR FOLLOW-UP PERIOD. THE OVERALL ANNUAL LOCAL POPULATION OF OLDER THAN 65 YEARS OLD IN THE "NON-FRAIL" CLASS IS SPECIFIED BETWEEN SQUARE BRACKETS.

	Observed events in a four-year follow-up (% ± 95% CI)			
Risk classes	**2012/2013** [59,264]	**2013/2014** [61,543]	**2014/2015** [59,499]	**2015/2016** [58,725]
Non at risk (0-10.8)	981 (2.2 ± 0.1)	1,779 (3.9 ± 0.2)	1,430 (3.2 ± 0.2)	1,309 (3.0 ± 0.2)
At risk (10.8-100)	3,837 (25.9 ± 0.7)	5,866 (37.7 ± 0.8)	4,663 (31.2 ± 0.7)	4,629 (31.1 ± 0.7)

performed to evaluate the goodness of fit. On the training set the AUROC was 0.8752 and Brier test returned a score of 0.0606; similar values were obtained on the test set: 0.8795 for the AUROC value and 0.0598 for the Brier score test. Value over 0.85 for AUROC curve identify an excellent prediction model and the lower is the Brier score, the better is the goodness of fit. Therefore, the two estimates demonstrated a good discrimination capability and calibration of the model.

The appropriateness of the model was further assessed by illustrating the observed outcome rates (i.e., the number of non-frail subjects becoming frail within a year) occurring in a four-year follow-up period from 2013 to 2016 on the overall annual local population. Table III shows the observed events in both classes with the related percentage distributions and 95% CI computed using Poisson regression. For each year, the overall annual local non-frail population older than 65 years old is specified between square brackets. The observed events are reported according to expected risk class identified by the *worsening risk* prediction model. For instance, in the "2012/2013" column we observed 981 non at risk and 3,837 at risk subjects becoming frail in 2013 according to the initially described *frailty risk* prediction model. All these subjects belonged to the "non-frail" class in 2012.
In each year, the percentage of the observed event for each class fell within the risk prediction range of the class. For instance, in 2013 among the subjects belonging to the "non-frail" class in 2012, the percentage of the observed event was 2.2% (within the range 0-10.8%), meaning that only 2.2% of the non-frail subjects labeled as "non at risk" turned into frail in 2013.

VIII. CONCLUSIONS

Ageing is becoming a meaningful challenge for many countries from social, financial and economic perspectives. Detecting frailty in elderly people represents a crucial research problem.

In this paper we proposed two frailty prediction models using a wide set of routinely collected data available from 12 socio-clinical databases. The models were built on the whole elderly population of the Municipality of Bologna and included clinical and socio-economic variables. The first model detected and categorized frailty according to the expected risk of emergency hospitalization or death within a year. Five classes with increasing *frailty risk* were identified and internal and external validations were performed demonstrating a good predictive ability. The second model assigned a *worsening risk* score to non-frail individuals according to their probability of becoming frail within a year. Similarly, an internal validation demonstrated the appropriateness of this model.

The strengths of our study include the possibility to guide appropriate planning of health resource utilization and develop patient-oriented preventive strategies. The use of routinely collected socio-clinical data reduced the potential risk of missing data and allowed to collect a wide variety of predictor variables including clinical and socio-economic aspects. Moreover, it represented a step forward to better meet a broader definition of frailty and greatly reduced the risk of referral and diagnostic biases.

The models might be applicable in a national and international setting with appropriate modifications and an extended tuning process, by investigating similar local socio-clinical and administrative databases. Furthermore, test-derived data and further clinical information (e.g., gait speed, grip strength, presence of tremor or mood disorders) can be included in the models, in order to improve the prediction ability.

APPENDIX A

Table IV summarizes all the variables provided by the 12 data sources. In Table V we provide the groups with similar pathologies created according to the *ICD-9-CM* taxonomy. In Tables VI and VII we show the odds ratios, *p*-values and 95% confidence intervals associating the socio-clinical variables with the outcomes included in both models.

ACKNOWLEDGMENT

The work was partially supported by the Italian Ministry of Education, Universities and Research, as a part of the OPportunities for active and healthy LONgevity (OPLON, SCN_00176) project (law n. 391/July 5, 2012, "Smart Cities and Communities and Social Innovation", http://attiministeriali.miur.it/anno-2012/luglio/dd-05072012.aspx). The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors thank Professor Beatriz De La Iglesia, University of East Anglia for her comments that greatly improved the manuscript.

TABLE IV
ALL THE VARIABLES PROVIDED BY THE 12 DATA SOURCES.

Data sources	Variables
Care Benefits Monitoring System	attendance allowance, date and place of birth, fiscal code, name and surname
Civil Registry	date and place of birth, deprivation index, education level, emigration date (if any), family income, fiscal code, household size, housing condition, level of education, marital status, name and surname, residence district
Death Registry	date and place of birth, date of death, fiscal code, name and surname
Emergency Department Registry	date and place of birth, fiscal code, name and surname, number of Emergency Department visits
Health Registry	address, age, date and place of birth, fiscal code, general practitioner, name and surname, nationality, residence district, sex
Hospital Discharge Record	date and place of birth, emergency hospitalization, fiscal code, hospitalizations with different discharge diagnosis code, name and surname, scheduled hospitalization and ICD-9-CM codes related to the following comorbidities: cancer, cerebrovascular diseases, chronic kidney disease, dementia, diabetes, heart failure, hypercholesterolemia, hypertension, liver disease, myocardial infarction, Parkinson's disease, respiratory diseases, surgical treatment for femoral fractures, thyroid diseases, vascular diseases
Hospital Pharmaceutical Registry	date and place of birth, fiscal code, name and surname, polypharmacy and ATC codes related to the following medications: cancer, diabetes, heart failure, myocardial infarction, Parkinson's disease, psychiatric disorders, respiratory diseases, thyroid diseases
Medical Exemption Certificates Registry	date and place of birth, fiscal code, name and surname and medical exemption certificates related to: arthritis, cancer, cerebrovascular diseases, chronic kidney disease, deaf and/or mute and/or blind, dementia, diabetes, gastric disease, heart failure, hypercholesterolemia, hypertension, liver disease, myocardial infarction, Parkinson's disease, physical disability, psychiatric disorders, respiratory diseases, vascular disease
Mental Health Department Registry	date and place of birth, fiscal code, name and surname of subjects diagnosed with mental diseases and regularly followed-up by Mental Health Departments
Out-of-hospital Assistance Registry	date and place of birth, fiscal code, name and surname, number of diagnostic tests or clinical exams performed in the out-of-hospital setting or in the Emergency Department
Public Home Healthcare Registry	date and place of birth, fiscal code, name and surname, duration, frequency and costs of home-based care
Territorial Pharmaceutical Registry	date and place of birth, fiscal code, name and surname, polypharmacy and ATC codes related to the following medications: cancer, diabetes, heart failure, myocardial infarction, Parkinson's disease, psychiatric disorders, respiratory diseases, thyroid diseases

TABLE V
SIMILAR COMORBIDITIES GROUPS.

Group name	Pathologies included
Cerebrovascular disease	subarachnoid, cerebral and intracranial hemorrhage, occlusion and stenosis of precerebral and cerebral arteries, transient cerebral ischemia, acute cerebral circulatory disorders and brain circulatory disorders
Dementia	senile and presenile organic psychotic conditions and Alzheimer's
Disability	all physical disabilities
Hypercholesterolemia	pure hypercholesterolemia and hyperglyceridemia and mixed hyperlipidemia
Hypertension	essential hypertension, hypertensive heart and chronic kidney disease, hypertensive nephropathy and secondary hypertension
Gastric disease	ulcerative colitis and Crohn disease
Liver disease	chronic hepatitis and cirrhosis
Psychiatric disorders	mental disorders, psychosis and anorexia
Respiratory diseases	acute and chronic bronchitis, emphysema, asthma, bronchiectasis and extrinsic allergic alveolitis
Thyroid diseases	simple and unspecified goiter, nontoxic nodular goiter, thyrotoxicosis with or without goiter and congenital hypothyroidism and other disorders of thyroid
Vascular disease	acute myocardial infarction, previous myocardial infarction, acute or subacute cardiac ischemia, angina pectoris and other forms of chronic ischemic heart disease

TABLE VI
ODDS RATIOS, *p*-VALUE AND 95% CONFIDENCE INTERVALS ASSOCIATING THE SOCIO-CLINICAL VARIABLES (IN BOLD) INCLUDED IN THE *FRAILTY RISK* PREDICTION MODEL.

Variables	**ORs**	***p*-values**	**95% CIs**	
Age (classes, ref. cat. 65-68)				
68-70	1.221	0.002	1.078	1.384
71-73	1.388	0.000	1.229	1.566
74-76	1.712	0.000	1.519	1.930
77-79	2.133	0.000	1.899	2.401
80-82	2.558	0.000	2.274	2.877
83-85	3.291	0.000	2.922	3.708
86-88	3.916	0.000	3.460	4.433
89-91	4.995	0.000	4.366	5.715
>91	6.395	0.000	5.524	7.402
Sex (ref. cat. Female)				
Male	1.372	0.000	1.300	1.448
Nationality (ref. cat. Foreign)				
Italian	1.004	0.967	0.841	1.198
Housing condition (ref. cat. Privately-owned)				
Renting	1.060	0.026	1.007	1.115
Other	0.964	0.489	0.868	1.070
Level of education (ref. cat. Higher)				
Primary	1.165	0.000	1.083	1.254
Secondary	1.094	0.023	1.012	1.182
Income (k€/year, ref. cat. >75k€/year)				
36-75	1.044	0.758	0.792	1.377
28-36	1.195	0.099	0.967	1.477
15-28	1.241	0.041	1.009	1.527
<15	1.444	0.001	1.172	1.780
Marital status (ref. cat. Married)				
Divorced	1.138	0.000	1.070	1.210
Single	1.224	0.000	1.117	1.342
Widowed	1.324	0.000	1.153	1.520
Deprivation Index (ref. cat. Very rich)				
Rich	1.010	0.801	0.934	1.093
Medium	1.071	0.081	0.992	1.156
Poor	1.093	0.027	1.010	1.184
Very poor	1.115	0.003	1.037	1.199
Hospitalizations (ref. cat. 0)				
1	1.241	0.000	1.164	1.324
2	1.356	0.000	1.239	1.484
>2	1.894	0.000	1.708	2.099
Emergency Room visits (ref. cat. 0)				
1	1.202	0.000	1.134	1.274
2	1.392	0.000	1.284	1.508
>2	1.562	0.000	1.432	1.703
Emergency hospitalization (ref. cat. No)				
Yes	1.980	0.000	1.664	2.356
Diagnostic test (ref. cat. No)				
Yes	1.431	0.000	1.199	1.707
Polypharmacy (>3 prescribed drugs, ref. cat. No)				
Yes	1.187	0.000	1.126	1.251
Benefits (ref. cat. No)				
Attendance allowance	1.281	0.018	1.044	1.572
Follow-up to MHD	1.373	0.000	1.160	1.624
Home-based care	1.549	0.000	1.450	1.654
Comorbidity (ref. cat. No)				
Cerebrovascular disease	1.189	0.001	1.078	1.311
Disability	1.264	0.000	1.175	1.360
Psychiatric disorders	1.308	0.000	1.226	1.395
Diabetes	1.379	0.000	1.292	1.472
Dementia	1.448	0.000	1.271	1.650
Chronic kidney disease	1.462	0.000	1.304	1.639
Respiratory diseases	1.488	0.000	1.351	1.638
Parkinson's disease	1.505	0.000	1.329	1.706
Liver disease	1.509	0.000	1.261	1.805
Cancer	1.517	0.000	1.424	1.616
Vascular disease	1.546	0.000	1.451	1.648
(Intercept)	0.021	0.000	0.017	0.026

TABLE VII
ODDS RATIOS, *p*-VALUE AND 95% CONFIDENCE INTERVALS ASSOCIATING THE SOCIO-CLINICAL VARIABLES (IN BOLD) INCLUDED IN THE *WORSENING RISK* PREDICTION MODEL.

Variables	**ORs**	***p*-values**	**95% CIs**	
Age (classes, ref. cat. 65-68)				
68-70	2.380	0.000	1.981	2.861
71-73	5.052	0.000	4.249	6.006
74-76	1.44×10^1	0.000	1.21×10^1	1.72×10^1
77-79	3.76×10^1	0.000	3.11×10^1	4.53×10^1
80-82	1.10×10^2	0.000	9.00×10^1	1.34×10^2
83-85	3.18×10^2	0.000	2.54×10^2	3.98×10^2
86-88	1.02×10^3	0.000	7.88×10^2	1.33×10^3
89-91	2.71×10^3	0.000	1.68×10^3	4.38×10^3
>91	8.73×10^3	0.000	5.35×10^2	1.42×10^5
Sex (ref. cat. Female)				
Male	4.518	0.000	4.146	4.923
Nationality (ref. cat. Foreign)				
Italian	0.765	0.174	0.519	1.126
Housing condition (ref. cat. Privately-owned)				
Renting	1.243	0.000	1.156	1.338
Other	0.942	0.422	0.813	1.091
Level of education (ref. cat. Higher)				
Primary	2.295	0.000	2.070	2.545
Secondary	1.594	0.000	1.434	1.772
Income (k€/year) (k€/year, ref. cat. >75k€/year)				
36-75	1.442	0.058	0.988	2.105
28-36	2.486	0.000	1.835	3.367
15-28	3.212	0.000	2.384	4.326
<15	6.554	0.000	4.840	8.874
Marital status (ref. cat. Married)				
Divorced	1.826	0.000	1.676	1.989
Single	2.440	0.000	2.145	2.775
Widowed	3.706	0.000	3.083	4.454
Deprivation Index (ref. cat. Very rich)				
Rich	1.109	0.069	0.992	1.239
Medium	1.334	0.000	1.199	1.484
Poor	1.450	0.000	1.292	1.625
Very poor	1.735	0.000	1.562	1.927
Hospitalizations (ref. cat. 0)				
1	1.704	0.000	1.531	1.897
2	2.928	0.000	2.423	3.539
>2	11.131	0.000	7.540	16.432
Emergency Room visits (ref. cat. 0)				
1	1.731	0.000	1.583	1.892
2	2.980	0.000	2.570	3.458
>2	5.603	0.000	4.719	6.653
Diagnostic test (ref. cat. No)				
Yes	2.476	0.008	1.265	4.847
Polypharmacy (>3 prescribed drugs, ref. cat. No)				
Yes	1.790	0.000	1.654	1.937
Benefits (ref. cat. No)				
Attendance allowance	2.821	0.002	1.474	5.399
Follow-up to MHD	4.065	0.000	3.133	5.274
Home-based care	9.403	0.000	7.453	11.864
Comorbidity (ref. cat. No)				
Cerebrovascular disease	3.136	0.000	2.495	3.942
Disability	3.386	0.000	2.965	3.868
Psychiatric disorders	3.669	0.000	3.273	4.112
Diabetes	4.746	0.000	4.267	5.279
Liver disease	6.422	0.000	4.764	8.655
Cancer	7.017	0.000	6.286	7.832
Respiratory diseases	8.496	0.000	6.641	10.870
Chronic kidney disease	8.874	0.000	6.291	12.517
Vascular disease	10.395	0.000	9.180	11.769
Parkinson's disease	10.863	0.000	8.531	13.834
Dementia	13.180	0.000	7.815	22.228
(Intercept)	0.000	0.000	0.000	0.000

REFERENCES

[1] L. Wang, S. Hu, G. Betis, and R. Ranjan, "A computing perspective on smart city [guest editorial]," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1337–1338, 2016.

[2] C. G. Cassandras, "Smart cities as cyber-physical social systems," *Engineering*, vol. 2, no. 2, pp. 156–158, 2016.

[3] Department of Economic and Social Affairs, "World population ageing 2015: Highlights," United Nations, Tech. Rep., 2015.

[4] F. Buckinx, Y. Rolland, J.-Y. Reginster, C. Ricour, J. Petermans, and O. Bruyère, "Burden of frailty in the elderly population: perspectives for a public health challenge," *Archives of Public Health*, vol. 73, no. 1, p. 1, 2015.

[5] T. A. Comans, N. M. Peel, R. E. Hubbard, A. D. Mulligan, L. C. Gray, and P. A. Scuffham, "The increase in healthcare costs associated with frailty in older people discharged to a post-acute transition care program," *Age and ageing*, p. afv196, 2016.

[6] T. Neuman, J. Cubanski, J. Huang, and A. Damico, "The rising cost of living longer: Analysis of Medicare spending by age for beneficiaries in traditional Medicare," *The Henry J. Kaiser Family Foundation*, 2015.

[7] K. E. Skouby, A. Kivimäki, L. Haukiputo, P. Lynggaard, and I. M. Windekilde, "Smart cities and the ageing population," in *The 32nd Meeting of WWRF*, 2014.

[8] E. Wang and J. Shi, "How smart city supports the travel of elderly care communities," in *Smart Cities Conference (ISC2), 2016 IEEE International*. IEEE, 2016, pp. 1–5.

[9] P. Paolini, N. Di Blas, S. Copelli, and F. Mercalli, "City4age: Smart cities for health prevention," in *Smart Cities Conference (ISC2), 2016 IEEE International*. IEEE, 2016, pp. 1–4.

[10] A. Hussain, R. Wenbi, A. L. da Silva, M. Nadher, and M. Mudhish, "Health and emergency-care platform for the elderly and disabled people in the smart city," *Journal of Systems and Software*, vol. 110, pp. 253–263, 2015.

[11] L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W. J. Kop, G. Burke *et al.*, "Frailty in older adults evidence for a phenotype," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 56, no. 3, pp. M146–M157, 2001.

[12] A. Clegg, J. Young, S. Iliffe, M. O. Rikkert, and K. Rockwood, "Frailty in elderly people," *The Lancet*, vol. 381, no. 9868, pp. 752–762, 2013.

[13] L. Rodriguez-Mañas and L. P. Fried, "Frailty in the clinical scenario," *The Lancet*, vol. 385, no. 9968, pp. e7–e9, 2015.

[14] S. L. Szanton, C. L. Seplaki, R. J. Thorpe, J. K. Allen, and L. P. Fried, "Socioeconomic status is associated with frailty: the Women's Health and Aging Studies," *Journal of epidemiology and community health*, vol. 64, no. 01, pp. 63–67, 2010.

[15] E. O. Hoogendijk, H. P. van Hout, M. W. Heymans, H. E. van der Horst, D. H. Frijters, M. I. B. van Groenou, D. J. Deeg, and M. Huisman, "Explaining the association between educational level and frailty in older adults: results from a 13-year longitudinal study in the Netherlands," *Annals of epidemiology*, vol. 24, no. 7, pp. 538–544, 2014.

[16] A. Clegg, C. Bates, J. Young, R. Ryan, L. Nichols, E. A. Teale, M. A. Mohammed, J. Parry, and T. Marshall, "Development and validation of an electronic frailty index using routine primary care electronic health record data," *Age and ageing*, vol. 45, no. 3, pp. 353–360, 2016.

[17] G. Turner, A. Clegg, A. Sayer, Avan, E. Burns, C. Beech, T. Dening, T. Gentry, J. Hindle, S. Iliffe, F. Martin, C. McAlpine, C. Nicholson, C. Patterson, J. Preston, and J. Young, "Fit for frailty. consensus best practice guidance for the care of older people living with frailty in community and outpatient settings," British Geriatrics Society in association with the Royal College of General Practitioners and Age UK, Tech. Rep., 2014.

[18] G. Turner and A. Clegg, "Best practice guidelines for the management of frailty: a British Geriatrics Society, Age UK and Royal College of General Practitioners report," *Age and ageing*, vol. 43, no. 6, pp. 744–747, 2014.

[19] D. Wennberg, M. Siegel, B. Darin, N. Filipova, R. Russell, L. Kenney, K. Steinort, T.-R. Park, G. Cakmakci, J. Dixon, N. Curry, and J. Billings, "Combined predictive model: final report and technical documentation," London: Department of Health, The Kings Fund, NYU, Health Dialogue, Tech. Rep., 2006.

[20] L. P. Fried, L. Ferrucci, J. Darer, J. D. Williamson, and G. Anderson, "Untangling the concepts of disability, frailty, and comorbidity: implications for improved targeting and care," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 59, no. 3, pp. M255–M263, 2004.

[21] D. Hamerman, "Toward an understanding of frailty," *Annals of internal medicine*, vol. 130, no. 11, pp. 945–950, 1999.

[22] P. Falasca, A. Berardo, and F. D. Tommaso, "Development and validation of predictive MoSaiCo (Modello Statistico Combinato) on emergency admissions: can it also identify patients at high risk of frailty?" *Annali dell'Istituto superiore di sanità*, vol. 47, no. 2, pp. 220–228, 2011.

[23] A. Kingston, K. Davies, J. Collerton, L. Robinson, R. Duncan, T. B. Kirkwood, and C. Jagger, "The enduring effect of education-socioeconomic differences in disability trajectories from age 85 years in the Newcastle 85+ study," *Archives of gerontology and geriatrics*, vol. 60, no. 3, pp. 405–411, 2015.

[24] F. Lally and P. Crome, "Understanding frailty," *Postgraduate medical journal*, vol. 83, no. 975, pp. 16–20, 2007.

[25] L. Calzà, D. Beltrami, G. Gagliardi, E. Ghidoni, N. Marcello, R. Rossini-Favretti, and F. Tamburini, "Should we screen for cognitive decline and dementia?" *Maturitas*, vol. 82, no. 1, pp. 28–35, 2015.

[26] D. H. Lee, K. J. Buth, B.-J. Martin, A. M. Yip, and G. M. Hirsch, "Frail patients are at increased risk for mortality and prolonged institutional care after cardiac surgery," *Circulation*, vol. 121, no. 8, pp. 973–978, 2010.

[27] J. Hippisley-Cox and C. Coupland, "Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score," *BMJ open*, vol. 3, no. 8, p. e003482, 2013.

[28] D. H. Kim and S. Schneeweiss, "Measuring frailty using claims data for pharmacoepidemiologic studies of mortality in older adults: evidence and recommendations," *Pharmacoepidemiology and drug safety*, vol. 23, no. 9, pp. 891–901, 2014.

[29] S. A. Sternberg, A. W. Schwartz, S. Karunananthan, H. Bergman, and A. Mark Clarfield, "The identification of frailty: a systematic literature review," *Journal of the American Geriatrics Society*, vol. 59, no. 11, pp. 2129–2138, 2011.

[30] P. Pandolfi, P. Marzaroli, M. Musti, E. Stivanello, and N. Collina, "Un modello statistico previsionale per misurare la fragilità." in *La fragilità degli anziani. Strategie, progetti, srtrumenti per invecchiare bene.*, G. Cavazza and C. Malvi, Eds. Maggioli Editore, 2014, ch. 2, pp. 27–36.

[31] W. Raghupathi and J. Tan, "Strategic it applications in health care," *Communications of the ACM*, vol. 45, no. 12, pp. 56–61, 2002.

[32] J. R. Schubart and J. S. Einbinder, "Evaluation of a data warehouse in an academic health sciences center," *International journal of medical informatics*, vol. 60, no. 3, pp. 319–333, 2000.

[33] J. S. Einbinder, K. W. Scully, R. D. Pates, J. R. Schubart, and R. E. Reynolds, "Case study: a data warehouse for an academic medical center," *Journal of Healthcare Information Management*, vol. 15, no. 2, pp. 165–176, 2001.

[34] J. L. Breault, C. R. Goodall, and P. J. Fos, "Data mining a diabetic data warehouse," *Artificial intelligence in medicine*, vol. 26, no. 1, pp. 37–54, 2002.

[35] M. F. Wisniewski, P. Kieszkowski, B. M. Zagorski, W. E. Trick, M. Sommers, and R. A. Weinstein, "Development of a clinical data warehouse for hospital infection control," *Journal of the American Medical Informatics Association*, vol. 10, no. 5, pp. 454–462, 2003.

[36] J. A. Lyman, K. Scully, and J. H. Harrison, "The development of health care data warehouses to support data mining," *Clinics in laboratory medicine*, vol. 28, no. 1, pp. 55–71, 2008.

[37] E. Roelofs, L. Persoon, S. Nijsten, W. Wiessler, A. Dekker, and P. Lambin, "Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial," *Radiotherapy and Oncology*, vol. 108, no. 1, pp. 174–179, 2013.

[38] M. De Mul, P. Alons, P. Van der Velde, I. Konings, J. Bakker, and J. Hazelzet, "Development of a clinical data warehouse from an intensive care clinical information system," *Computer methods and programs in biomedicine*, vol. 105, no. 1, pp. 22–30, 2012.

[39] M. D. Krasowski, A. Schriever, G. Mathur, J. L. Blau, S. L. Stauffer, and B. A. Ford, "Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research," *Journal of pathology informatics*, vol. 6, 2015.

[40] K. L. Brigham, "Predictive health: the imminent revolution in health care," *Journal of the American Geriatrics Society*, vol. 58, no. s2, pp. S298–S302, 2010.

[41] N. Peek, C. Combi, R. Marin, and R. Bellazzi, "Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes," *Artificial intelligence in medicine*, vol. 65, no. 1, pp. 61–73, 2015.

[42] C. K. Reddy, C. C. Aggarwal, C. K. Reddy, and C. C. Aggarwal, "An introduction to healthcare data analytics," in *Healthcare Data Analytics*. Chapman and Hall/CRC, 2015, pp. 1–18.

[43] H. C. Koh, G. Tan *et al.*, "Data mining applications in healthcare," *Journal of healthcare information management*, vol. 19, no. 2, p. 65, 2011.

[44] B. Milovic and M. Milovic, "Prediction and decision making in health care using data mining," *Kuwait Chapter of the Arabian Journal of Business and Management Review*, vol. 1, no. 12, p. 126, 2012.

[45] P. Johnson, L. Vandewater, W. Wilson, P. Maruff, G. Savage, P. Graham, L. S. Macaulay, K. A. Ellis, C. Szoeke, R. N. Martins *et al.*, "Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease," *BMC bioinformatics*, vol. 15, no. 16, p. 1, 2014.

[46] J. Zhou, J. Sun, Y. Liu, J. Hu, and J. Ye, "Patient risk prediction model via top-k stability selection," in *SIAM conference on data mining. SIAM*. SIAM, 2013.

[47] F. Babič, L. Majnarić, A. Lukáčová, J. Paralič, and A. Holzinger, "On patient's characteristics extraction for metabolic syndrome diagnosis: Predictive modelling based on machine learning," in *International Conference on Information Technology in Bio-and Medical Informatics*. Springer, 2014, pp. 118–132.

[48] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," in *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on*. IEEE, 2012, pp. 173–177.

[49] J. S. Mathias, A. Agrawal, J. Feinglass, A. J. Cooper, D. W. Baker, and A. Choudhary, "Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data," *Journal of the American Medical Informatics Association*, vol. 20, no. e1, pp. e118–e124, 2013.

[50] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[51] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using twitter," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1568–1576.

[52] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *International journal of medical informatics*, vol. 77, no. 2, pp. 81–97, 2008.

[53] C. K. Reddy and Y. Li, "A review of clinical prediction models," in *Healthcare Data Analytics*. Chapman and Hall/CRC, 2015, pp. 343–378.

[54] A. Vaisman and E. Zimányi, *Data Warehouse Systems*. Springer, 2014.

[55] M. Dallachiesa, A. Ebaid, A. Eldawy, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang, "NADEEF: a commodity data cleaning system," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013, pp. 541–552.

[56] S. G. Rizzo, D. Montesi, A. Fabbri, and G. Marchesini, "Icd code retrieval: Novel approach for assisted disease classification," in *International Conference on Data Integration in the Life Sciences*. Springer, 2015, pp. 147–161.

[57] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304.

[58] J. F. J. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 7th ed. Pearson, 2014.

[59] H. Khedmat, G.-R. Karami, V. Pourfarziani, S. Assari, M. Rezailashkajani, and M. Naghizadeh, "A logistic regression model for predicting health-related quality of life in kidney transplant recipients," in *Transplantation proceedings*, vol. 39, no. 4. Elsevier, 2007, pp. 917–922.

[60] P. Bosi, M. Lorenzini, A. Scagliarini, F. Paltrinieri, C. Lambertini, and F. Bertoni, "Il sistema informativo socio-sanitario e il supporto all'attività di programmazione," Centro di Analisi delle Politiche Pubbliche, Tech. Rep. CAPPaper n. 100, 2012.

[61] R. Fabrizio and E. Verdini, "Flusso informativo: Assistenza residenziale e semiresidenziale per anziani o persone non autosufficienti in condizioni di cronicit e/o relativa stabilizzazione delle condizioni cliniche," Giunta Regione Emilia-Romagna, Tech. Rep. PG.2010.0041963, 2010.

[62] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[63] G. Ridgeway, "Generalized boosted models: A guide to the gbm package," *Update*, vol. 1, no. 1, p. 2007, 2007.

[64] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[65] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

Flavio Bertini received the Master Degree Summa Cum Laude in Computer Science in 2011 and the Ph.D. in Computer Science from the University of Bologna in 2015. In 2013, he was a visiting researcher at the Inria Bordeaux - Sud-Ouest. Currently, he is a research fellow with the SmartData Research Group. His research interests focus on online fingerprinting, predictive models in healthcare and large-scale data analytics.

Giacomo Bergami received the Master Degree Summa Cum Laude in Computer Science in 2014. In 2017 he was a visiting student at University of Leipzig, Germany, where he worked on the Gradoop Project. Currently he is working with prof. Danilo Montesi as a Ph.D. student and a member of the SmartData Research Group. His research interests focus on data mining, graph database management systems and data uncertainty.

Danilo Montesi is a full professor of database and information systems at the Department of Computer Science and Engineering of the University of Bologna since 2005. He has held visiting researcher positions at the Department of Computer Systems and Telematics, University of Trondheim, the Imperial College of London, the Department of Computing, Purdue University, the Rutherford Appleton Laboratory (UK) and the University of Lisboa. He is the founder of SmartData Research Group.

Giacomo Veronese is currently enrolled into the Emergency Medicine Residency Program at the University of Milano-Bicocca (Niguarda Ca Granda Hospital, Milan, Italy). During his training, while focusing the attention towards the clinical aspects of emergency medicine, he has been enrolled as a research fellow in several universities, mostly focusing on data collection, statistics, clinical epidemiology and research methods.

Giulio Marchesini is full professor of Dietetics at University of Bologna since 2012. From 2005 to 2012, he was full professor of Internal Medicine. Since 2006, he is head of Clinical Dietetics at S. Orsola-Malpighi University Hospital, Bologna. His principal interests are in nutrition in patients with diabetes and with non alcoholic fatty liver disease; disease management of the metabolic syndrome; epidemiology, treatment and outcome assessment. He is member of several international scientific committees.

Paolo Pandolfi is specialised in Hygiene, Preventive Medicine and Public Health and he is currently head of the Epidemiology and Health Promotion Unit, Department of Public Health, AUSL Bologna. He has been a contract professor for the School of Medicine, University of Bologna. His principal interests cover epidemiology, health promotion and risk prevention. He publishes scientific works on preventive medicine and epidemiology.