World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:16, No:4, 2022

Using Statistical Significance and Prediction to Test Long/Short Term Public Services and Patients Cohorts: A Case Study in Scotland

Sotirios Raptis

Abstract—Health and Social care (HSc) services planning and scheduling are facing unprecedented challenges, due to the pandemic pressure and also suffer from unplanned spending that is negatively impacted by the global financial crisis. Data-driven approaches can help to improve policies, plan and design services provision schedules using algorithms that assist healthcare managers to face unexpected demands using fewer resources. The paper discusses services packing using statistical significance tests and machine learning (ML) to evaluate demands similarity and coupling. This is achieved by predicting the range of the demand (class) using ML methods such as Classification and Regression Trees (CART), Random Forests (RF), and Logistic Regression (LGR). The significance tests Chi-Squared and Student's test are used on data over a 39 years span for which data exist for services delivered in Scotland. The demands are associated using probabilities and are parts of statistical hypotheses. These hypotheses, as their NULL part, assume that the target demand is statistically dependent on other services' demands. This linking is checked using the data. In addition, ML methods are used to linearly predict the above target demands from the statistically found associations and extend the linear dependence of the target's demand to independent demands forming, thus, groups of services. Statistical tests confirmed ML coupling and made the prediction statistically meaningful and proved that a target service can be matched reliably to other services while ML showed that such marked relationships can also be linear ones. Zero padding was used for missing years records and illustrated better such relationships both for limited years and for the entire span offering long-term data visualizations while limited years periods explained how well patients numbers can be related in short periods of time or that they can change over time as opposed to behaviours across more years. The prediction performance of the associations were measured using metrics such as Receiver Operating Characteristic (ROC), Area Under Curve (AUC) and Accuracy (ACC) as well as the statistical tests Chi-Squared and Student. Co-plots and comparison tables for the RF, CART, and LGR methods as well as the p-value from tests and Information Exchange (IE/MIE) measures are provided showing the relative performance of ML methods and of the statistical tests as well as the behaviour using different learning ratios. The impact of k-neighbours classification (k-NN), Cross-Correlation (CC) and C-Means (CM) first groupings was also studied over limited years and for the entire span. It was found that CART was generally behind RF and LGR but in some interesting cases, LGR reached an AUC = 0 falling below CART, while the ACC was as high as 0.912 showing that ML methods can be confused by zero-padding or by data's irregularities or by the outliers. On average, 3 linear predictors were sufficient, LGR was found competing well RF and CART followed with the same performance at higher learning ratios. Services were packed only when a significance level (p-value) of their association coefficient was more than 0.05. Social factors relationships were observed between home care services and treatment of old people, low birth weights, alcoholism, drug abuse, and emergency admissions. The work found that different HSc services can be well packed as plans of limited duration, across various services sectors, learning configurations, as confirmed by using statistical hypotheses.

Keywords—Class, cohorts, data frames, grouping, prediction, probabilities, services.

Mr. Raptis was with the School of Design and Informatics, Abertay University, Dundee, Scotland,Bell Street, Dundee, DD1 1HG and with (e-mail: sotiris.raptis.jb@gmail.com).

I. Introduction

THE current work aims at predicting H&Sc services usage by associating them in clusters of services using ML methods. The services data were public data from the National Health Services Scotland (NHSS), Public Health Services (PHS) web site [1] posted by June, 2019. The services are statistically related using the hypothesis tests $\chi^2$and Student (t-test) while the plausibility obtained was tested using the information exchange and the statistical significance level (p-value). The prediction of each demand was based on a group's behaviour rather than on a service's past attendances . The paper argues that statistics and ML can help to investigate statistical relationships between the services in order to study them better and to improve the allocation of healthcare (HC) resources to them so that the patients can benefit more. This can lead to less spending of resources on specialities, on man-hours, on IT, on medical equipment, etc., when they can be predicted. Then the services can be merged and cost less when they are attended by larger numbers of patients . The data to analyse are the year series of the services' parameters that define how they were delivered to the public. This work builds on works in which the methods discussed were used separately as in [2] in which CMand hierarchical clustering are used to offer an initial number of clusters . Then the features are sorted by importance, then their clusters are split or merged using homogeneity criteria. As a result of the splitting/merging the number of clusters changes. The prediction of the social parameters is also suggested in [3] where ML links clinical parameters (for example body mass index) or biological signals (for example blood pressure) to social status data such as income, education, etc.Social factors, as for example those discussed here, are also used in *'AlcoholYoung '* as main predictors of clinical or biological factors. They are also proved in the present work to be statistically connected to the factor *'Low birth weights.Value '* as a single service (target).

The long-term histograms over a period of 39 years of the 21 services data frames were taken resulting in 110-year series that are tracked. The factors' names and the years (called

dates groups (DGs)) have been originally recorded in by PHS are shown in Table I. The DGs contain non-zero records and zero-padding was applied to fill in missing years demands. Tables II and III give a detailed picture of the settings in which the services are offered to the public including their acronyms (short names). The short names will be used in this work and are the 1^{st}part of the services names' triplets. The triplets comprise a service's acronym as their 1^{st}part, then the orders (or the IDs) of the attributes follow as a 2^{nd}part, and as a 3^{d}part the attributes levels follow. This naming convention helps to refer to all aspects of the services as well as to encode the many and complex relationshipsfound between them. The naming convention for the attributes and their levels was built using Tables II and III. For example *'BMIDistribution.2.1 '* would stand for *'Body Mass Index-Distribution.Gender.All '* while *'HCServices.Value '* would have no specific attributes and would stand for the number of patients that are recorded for that factor while *'HCProvision.26.2 '* would represent HC provision for Adults with Learning Disabilities.

The statistical tests provide support (as a p-value) or the likelihood of them being in a group. Using these tests the services can be compared and thus can be offered as a plan at a lower operational cost. The tests assess whether the services are likely co-attended based on the information they exchange. A low p-value (< 0.05) shows dependence. One can then create a cheaper plan by dependent demands. The low exchange shows likely co-attendance which again shows likely membership to the same cohort. ML is used to see if one can safely (likely) predict one demand from the group/cohort it belongs to using as many as possible colinearities (other linearly related demands) that can be used as its predictors. The services are first linked using groupings from CM, or IE as well as from CC in different settings. LGR was found well competing while RF and CART followed both with the same performance but at higher learning ratios. Factors relationships between home care services and treatment of old people, or low birth weights, alcoholism, drug abuse, and emergency admissions were found.

Zero padding offered long-term data visualization while the attendances' relationships changed across DGs and over all years. The significance tests confirmed parts of the linearly derived relationships, that is, one-to-one relationshipsbetween a target and a single predictor.

The PHS data were read as data framesusing R that contained all the parameters that defined a service. The prediction and the statistical significance are helpful in mining services sequences or patterns of patients' behaviour using sequence optimization methods. These are used in [4] to mine stored processes patterns. In process mining applications the degree of similaritycan be a statistical hypothesis that can hold or not. This can be established using the p-value or the amount of information exchanged between these services. The similarity can be made more specific and expand the ways a relationship between data is considered. For example the similarity can mean a correlation of values as in the case of CC or it may mean a dependence on either a linear one or a statistical one. For example, such a relationshipis studied in [5] with respect to the dependence of the survival rate for patients in the ICUs on the degree of their blood oxygenation (PCO2). This relationship means that the target and the independent variables can be interpreted as a linear dependence between them that is governed by the high p-value ($p = 0.85$) of the significance test. This is a very high p-value that does not allow to eliminate the alternative (not NULL) hypothesis of the dependence of them. Another known way to combine regressionand hypothesis testing is used in [6] that checks the plausibility of the linear coefficients . Here, the reliable coefficients had a p-value of less than 0.05.

II. Information Exchange between histograms to Group H&Sc factors

CC can indicate a point-to-point similarity measure but it does not offer a formal way to check the validity of the similarity. In this work, the average CC was used as a criterion to include services in the same class. CC is based on the dot product of individual values vectors. When two vectors are not well linearly related (their LR has high residual errors) it is unlikely they have a high CC. On the other hand, IE does not account for single years attendances similarities in specific years. Thus, CC offers a more granular picture of the variation over years that is suitable for exact pattern matching whereas IE looks into the broader picture of how often the H&Sc services are similarly attended.

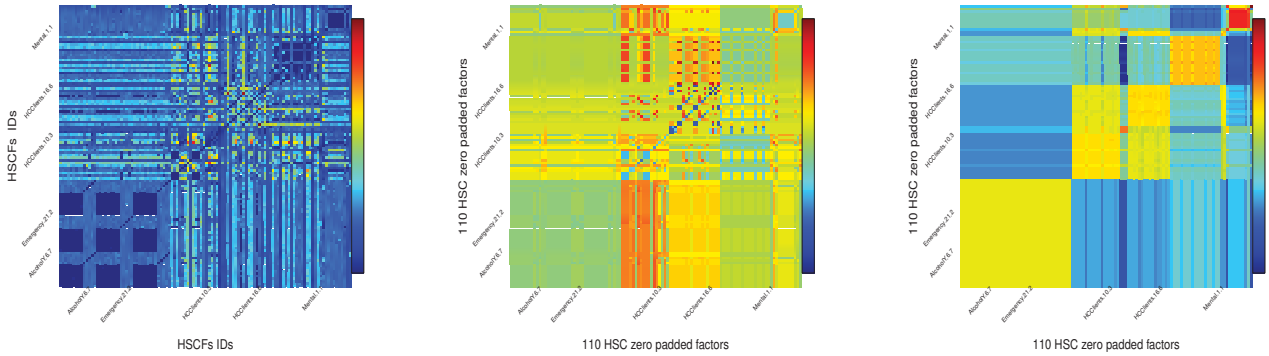
The IE was computed on all pairs of histograms of the H&Sc factors involved over the 39 yearsspan due to the limitations that DGs imposed. IE is also referred to as Mutual IE (MIE) and is in line with the concept of bins (values intervals) that are used to build histograms. This is because these can be more easily matched between H&Sc factors rather than when taking single values. The bins best describe categorical variables whose levels are taken as distinct intervals of values.

III. CC and MIE

CC was used to check pair-wise factors similarities leading to classificationwhen more H&Sc factors are found well cross-correlated. CC varies more than MIE does since it compares single values. MIE is more suitable for services and pathways grouping and gives a broader picture because it is less sensitive to single values as in an interval I, $a \in I = [b, c]$. It only takes into account the intervals Is. For two factors year series $TS_{i,j} i, j \in [1, 110]$ the cross-correlation is: $CC_{i,j} = \frac{COV(TS_i, TS_j)}{STD(TS_i) * STD(TS_J)}$

The most correlated factors were found in the same DGs. Some indicative well cross-correlated H&Sc factors are listed in Table VI. The table lists mainly H&Sc factors in different HSc groups.

To compare the attendances' frequencies and find out how rarely the services were jointly taken MIE was used as well. MIE is used to find redundant biomedical data or the best ones as explained in [7]. MIE is based on the IE that measures the likelihood of how often two data streams symbols occur together. High MIE for a pair of factors attendances $x \in HSC1$ and $y \in HSC2$ shows that their combination is



(a) (b) (c)

Fig. 1 (a) χ^2-based, (b) MIE-based, (c) CC-based match of the 110 series. 5 sampled factors

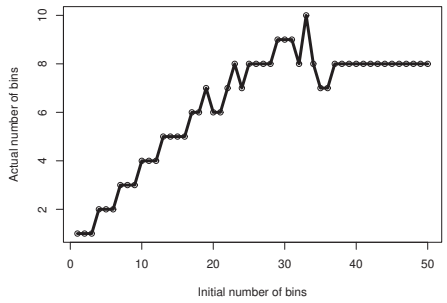


Fig. 2 The actual number of bins (truly attended) vs. the initial/no of bins suggests a number of bins around 9

not very likely. When this combination occurs it bears a lot of value. The IE is defined as:

$$IE(x,y) = log(\frac{1}{p(x,y)}) \quad (1)$$

The more unlikely is that two factors are co-attended in same years the higher their MIEis. MIE was computed on pairs of H&Sc factors attendances intervals and on common years. The probability used in (1) can be non-zero for pairs of intervals instead of single attendances that have nearly zero occurrence probability, i.e. , $a \in [b,c] \rightarrow p(a) = 0$. The inverse of it $[\frac{1}{p(x,y)}]$ shows how rare is this event. It is taken in its logarithmic form which can approach $-\infty = log(1/p[\rightarrow 0])$ if a single joint attendance were considered that can have a very low probability.

In [8] MIE is used to see how independent two clustering algorithms are when applied onto the same data. This is achieved by comparing the similarity or the dissimilarity of them. In the present work, MIE was used to compare how similar or dissimilar the attendances of two H&Sc factors were over the years.

To better apply MIE, the pairs of services' attendances were considered. The attendsnces' values (demands) were 1^{st} allocated into 9 equally large demand intervals and then the intervals assigned formed the pairs. This arrangement gave non-zero MIE. This made it possible to account from very low to very high attendances up to the maximum attendance that was found across factors. Among the most attended H&Sc factors some indicative were: *'HCDisability . Care Home Sector . Voluntary Sector '*(1444050), *'HCDisability . Main Client Group . in Care Home Adults with Mental Health Problems '*(1302029), *'HCDisability. Main Client Group . in Care Home Adults with Learning Disabilities '*(3092400). The least attended were *'LowBirthweight . Limiting long term physical or Mental Health Condition . Limiting condition '*(70), *'HCDisability . Main Client Group . In Care Home Adults with Learning Disabilities '*(5). Then, each factor's attendance was assigned a MIE interval from 1 to 9. Then, MIE was computed for all pairs of H&Sc factors as discussed in Section III. H&Sc factors that exchanged a low amount of information (low MIE or high co-attendance) were in the HSc group *'Deprivation by smoking behavior '*(SmokingYoung), with parameters age and gender, *'Smoking prevalence among 13 and 15 year old's in Scotland '*, etc.. Among the factors that exchanged high information (high MIE or low co-attendance) were: *'AlcoholStats. Age . 13 '*, *'GPWorkforce. Weight Category . Epidemiological-Obese '*, *'HCDisability . Weight Category . Clinical-Obese '*, *'HCDisability . HCClientsGroup . Learning Disability '*, *'GPWorkforce. Weight Category . Clinical-SeverelyObese '*. Low MIE means that the factors are likely to be co-attended. Indicative cases of low MIE and high co-attendance are given in Table V. On the other hand, high MIE factors are *'GPWorkforce .Gender.Male '*, *'HCClientsGroup . Gender . Female '*, *'HCClientsGroup . Gender . Male '*, *'HCClientsGroup . Gender. All '*, etc.. That means these are not often co-attended and the chance that they have similar attendances is not high. This also means that they have low CC. At this point it is interesting to observe the Figs. 1b and 1c that show how low MIE values can correspond to high CC values.

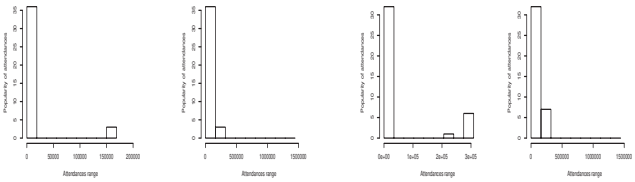
MIE was also computed between any two DF's. This was done by averaging MIE's over all single factors pairs

per H&Sc or per data frame. That means that if $HSCF1$ is a factor whose year series is denoted by $TS_{HSCF1} = [v_{1,1}, ...v_{1,t=n}]^T, n = 39$ and for $HSCF2$ the year series is $TS_{HSCF2} = [v_{2,1}, ...v_{2,t=n}]^T$ then the average MIE between the two H&Sc groups (as opposed to factors) is $MIE(1,2) = MIE(TS_1, TS_2) = 1/n * \sum_i (MIE(v_{1,i}, v_{2,i}))_{i=1}^{i=n}$, while $MIE(v_{1,i}, v_{2,i}) = p(v_{1,i}, v_{2,i}) * log(1/(p(v_{1,i}, v_{2,i})))$ and $p(v_{1,i}, v_{2,i})$ is the joint probability of pairing the two levels $v_{1,i}$ and $v_{2,i}$ from the two data streams HSCF1 and HSCF2.

Let us now assume that two HSc groups HSC1 and HSC2 have k and m numbers of factors, i.e., combinations of {attributes, levels} each. Then, we take all the existing combinations of their levels from the PHS data, i.e., all the rows where they are found together. Then, we can create the possible levels IDs combinations for the two attributes and we denote them as $IDs_{HSC1} = [1, k]$ and as $IDs_{HSC2} = [1, m]$. Then, the attributes combinations will be: $CMB_{IDs_{HSC_1}, IDs_{HSC_2}} = bn_{un_{1,i_k}, un_{2,j_m}} * \{i_k, j_m\}$. The indices un_{1,i_k} and un_{2,j_m} stand for the $i_k - th$ and $j_m - th$ unique levels of the two attributes when all the unique levels for attribute 1 are k and for attribute 2 are m. The quantity $bn_{(x,y)}$ is a binary flag that indicates that the two levels x and y can be observed together and are in the same row when the attributes (columns) 1 and 2 are examined. We can only take the existing combinations of levels for two attributes when computing the MIE. Each of these combinations includes the attributes and their levels from both HScs, i.e., $\{\{i_1, j_1\}, \{i_2, j_2\}\} \ni i_k, j_k \in IDs_{HSC_k} \ni k \in [1, 2]$. The (i_1, i_2) pair corresponds to a pair and has its own MIE, $MIE(i_1, i_2) = MIE(TS_1, TS_2)$. Then, we take the average of the $MIE(i_1, i_2)$'s to define the $MIE(HSC1, HSC2)$. For the 110 H&Sc factors and the entire observations period, the normalized MIE is shown in Fig. 1b. That means the MIE finally computed was an $N(\mu = 0, \sigma = 1)$ process. The range of the normalized MIE was $MIE \in [-5.914, 4.551]$ so that it can be compared easily with the normalized CC. Generally, a low MIE pair does not exactly but roughly correspond to a high CC pair as it is the case of predicting *'HCDisability. 26 . 3 '* (MIE=-0.596, CC=0.902). Average MIE values are also found together with average CC values as in the pair (MIE=0.103, CC=0.254). Same HSC group factors such as *'GPBoard . 8 . 2 '* and *'GPBoard . 9 . 3 '* do not differ much in how they are related to the target *'HCDisability. 26 . 3 '*. In this case, this means that the H&Sc factor "HCclients" can help in predicting the necessary HC workers such as the factor *'GPs '*.

The histograms for the full span of 39 yearswere computed taking different attendances resolutions to partition their values range. A number of bins around 9 worked well and divided the full attendances intervalinto bins that could be compared. No specific cut-points were used as limits for the bins and this allowed unbiased and homogeneous histograms with common ranges in the bins.

As discussed in [9] the number of bins is a trade-off of the statistical parameters (variance, values range,...) as well as the likelihood of being attended. A final number of bins was adopted as common for all data. This was found after different resolutions (initial numbers of bins) were tried. The



a b c d

Fig. 3 Normal and rescaled histograms for H&Sc factors (a-b) *'GPWorkforce.Weight category epidemiological . Underweight '*, (c-d) *'HCClientsGroup . Weight category clinical . Overweight '*

one that gave the minimal number of empty bins (ranges of attendances never observed) and at the same time seemed to be a stability point (local maximum) for the final number of bins was kept. An initial number of bins was chosen from the interval (1:50) and then the actual number was computed considering only the non-empty ones and how many attendances were well represented by a candidate-final number of bins. Then the attendance (popularity) of the factors to each candidate number of non-empty bins was derived. This was a function of both the variance of the factor while the number of bins that worked and gave a local maximum of 9 that also produced practically comparable histograms that spanned more bins. A diagram of the relationship between initial and final histograms resolutions is given in Fig. 2. The number of bins can guide the quality of the predictionas discussed in [10]. This is also observed here when a very coarse histogram defines less and large classes, it has fewer bins and offers a poor prediction while a more granular histogram has more bins and may result in many zero probabilities. That means that when the χ^2test was applied to histograms with many zeros (very small bins rarely busy) it would yield very low p-values and in turn, would make all factors independent. This is a biased decision though.

IV. Grouping Year Series Based on χ^2

The χ^2test was applied to infer the dependence or the independence between pairs of H&Sc factors leading to grouping or to not grouping them. χ^2 was applied to attendances histograms as discussed in Section II and is another expression of the probability that an attendances' interval is occupied. The χ^2test examines how similar are two attendances intervals(represented as bins) while the IE examines the probability of them occurring on the same year. Both methods can imply services grouping and are statistical methods that can be used as a complement to pure ML methods discussed to test services dependence as a statistical hypothesis. The dependence likely enables classification and prediction depending on the $p - value$ derived from the χ^2test. The low p-values define the plausibility of the NULL hypothesis that suggests the independence of two H&Sc factors [11]. Thus, these can be part of an H&Sc plan in the sense that the policymaker can predict the demand using cohort information.

Before applying $\chi^2$the re-scaled versions of the histograms $(h_{scaled,ts_i}, i \in [1, 110])$ of the zero-padded year series

TABLE I
H&SC ACRONYMS AND DATES GROUPS

Full Description	Acronyms	DateCode	DG
1 Alcohol use ever among young people (SALSUS)	AlcoholYoung	2002-2015	1
2 Emergency Admissions	EmergencyAdmissions	1998 - 2010	1
3 Headcount of GP workforce by Health Board and Local Authority	HeadcountGP	2008-2019	2
4 Number of home care clients by care type or disability(per 1,000 population)	HCDisability	2005-2009	1^{*2}
5 Home Care Client Living Arrang.	HCArrangements	2005 - 2009	2
6 Intensive Home Care	IntensiveHC	2002 - 2011	1
7 Places in single rooms in Care Homes	SingleRooms	2007-2017	2
8 Drug Related Hospital Discharge	DrugDischarges	1996-2018	2
9 Home Care Services	HCServices	2005-2009	1
10 Number[percent], low birthweight ($<$ 2500g) babies (single births)	Lowbirthweight	2000-2019	2
11 Mental well-being by tenure, household type, age, sex and disability	MentalDisability	2014-2017	4^{*3}
12 The number of GP's (GP surgeries) by area as at 1 October	GPWorkforce	2007-2019	2
13 Number of Homes Homes(by type of provision)	HCProvision	2007-2017	2
14 Occupancy rate in care homes by type of provision	Occupancy	2007-2017	2
15 Places in Care Homes with En-Suite Facilities	En-Suite	2007-2017	2
16 Primary 1 BMI Distribution	BMIDistribution	2001-2019	3
17 Smoking prevalence among 13 and 15-year olds in Scotland	SmokingYoung	2002-2015	2
18 Delayed discharges monthly census	DDMonth	2016-2020	4
19 Primary 1 Children Body Mass Index - Epidemiological	BMIChildren	2001-2019	4
20. Alcohol Related Hospital Statistics(updated after 6.2019)	AlcoholStats	1981-2019	5
21. Health Care Clients Group	HCClientsGroup	2016-2020	4

DG 1(2004-2010), 2(2010-2016),3(2012-2019),4(2014-2019),5(1981-2019) ^{*x} x is the number of padded years before inclusion to group. The "DateCode" are the years that are actually used to in the calculations and may differ from the original span

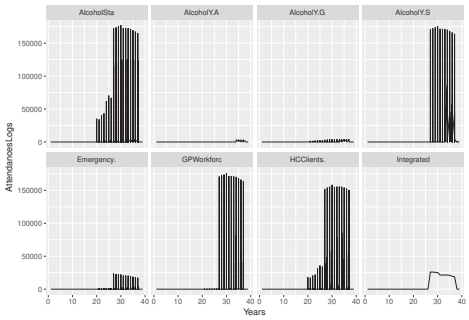


Fig. 4 Representative H&Sc factors plots that show the similarityand the variety of the attendances . The H&Sc factors are from top left to bottom right: (1) AlcoholStats.Age.13, (2) AlcoholY.Gender.All, (3) AlcoholY.Age.All, (4) AlcoholY.Smoking.Behaviour.Non-smoker, (5) Emergency.Age.All, (6) GPWorkforce.Gender.All (7) HCClients.Group.Gender.All, (8) IntegratedHC.Services.Value

were taken to have a common scale, such that all histograms are defined over the same global interval, ($I_g = [min(v_{1,1}, ..., v_{110,39}), max(v_{1,1}, ..., v_{110,39})$), where $v_{i,j} = HSCF_i(year = j)$. As discussed a single H&Sc factor is represented by the year series $TS_i = \{v_{i,1}, ..., v_{i,39}\}, i \in [1, 110]$ of single years attendances.

To match using χ^2 the single H&Sc factors we need to account for the entire spectrum of the individual attendances and not only the individual attendances per factor. As discussed the specific parameters under which they can be delivered are defined by the columns (attributes) in the PHS data. These attributes refer to the specific character of a service or to the patients' circumstances (example age or gender) who take it. This new interval is $I_g = [0, 176944]$. Hence, each bin contains $176944/9 \approx 19660$ patient visits.

Representative re-scaled histograms are shown in Figs. 3a-3c. The optimal number of bins means that before and after that specific number, the number of bins that were attended by different H&Sc factors was falling. A single factor's histogram has a set of bins that represents ranges of counts, over the span of 39 years for this. A usual number is in the range $(5, 50)$. Then the year series were compared in terms of their re-scaled histograms. The formula is given in (2):

$$\chi^2(k, l) = \sum_{j=1}^{j=9} \frac{(h_{scaled,k}(j) - h_{scaled,l}(j))^2}{h_{scaled,l}(j)} \qquad (2)$$

where $h_{scaled,k}(j)$ is the j^{th} bin of the k^{th} histogram. One can see that the attendances are not easily comparable in the middle range. The χ^2test was used with the assumption that one factor's attendance, that is, the k^{th}'s in (2) are the observed data and that the 2^{nd}factor, that is, the l^{th}s are the expected data.

Let us assume that all year series values (all possible demands values across the services) lay in the interval I_g. Then the separate year series ranges for each i, service ($[min_{ts_i}, max_{ts_i}], i \in [1, 110]$) occupy one or more bins (intervals) in this overall interval. The number of bins chosen yielded dependence across the re-scaled histograms and allowed all single observations to be considered. For example, a bin containing the interval $[0, 500]$ for some H&Sc factor for an overall range of values of $[0, 100000]$ would give many bins sized of size 500 while this H&Sc factor would only occupy one bin and would have zeros in the rest of the bins. This pattern would not be well correlated with another factor's attendances with more than 600 patients whose bin would be empty (although very similar) and would not fit in the bin of ($[0, 500]$). The remaining 100 patients (600-500) would need to be represented by another bin, i.e. ($[501, 1000]$). On the other hand, a single bin of ($[0, 100000]$) would accommodate many quite different histograms in a single interval and would not provide much insights for classification since all histograms would be in the same class. For a given resolution (number of bins) the dependence likelihood depends on the $(p - value)$ that is computed and it is shown in Fig. 1a. The connections (H&Sc factors pairs) laying onto or around the diagonal are the lowest (thus connected). This is expected since a histogram is mapped to itself. Thus, we seek high p-values that may be indicative of connections. Indeed, the p-values observed in Table VII are at the edge ($p - value = 0.05$) at which the NULL is valid and the interesting high p-values ($p > 0.05$) are marked with boldfaced numbers. The H&Sc factors IDs(i.e. , x/y-axis labels) in this figure are the same used in Tables III and II. The most correlated H&Sc factors using the χ^2 metrics are presented in Table VII by tracking the H&Sc factors that are well connected (low p-value) to a given H&Sc factor (row or column) in Fig. 1. The χ^2test can be used as a classifier by putting together (in the same class) the commonly attended H&Sc factors after 2000. Most services have records after that year. Indicative plots that show the similarity of the year series are given in Fig. 4. For example it is easy to see how the 1^{st},4^{th}, 6^{th} and 7^{th} resemble and then the 1^{st} and 2^{nd} are also similarly attended. Generally speaking, same H&Sc data frames have equal chances to be co-attended. This can be seen in the groups of H&Sc factors that are related to patients with disabilities as shown in Table IV and specifically for patterns with low MIE. Such high CC can also be observed in Table VI. Unlike CC, χ^2 is not based on the individual values of attendances but rather on how often these are taken.

TABLE II
LEVELS PER ATTRIBUTE USED IN THE SHORT NAMES [TRIPLETS] OF THE H&SC FACTORS

1:**Age**{"13" , "15", "All"}, **2**:**Gender**{"All", "Female", "Male" }, **3**: **Smoking Behaviour**{"Non-smoker", "Occasional smoker", "Regular smoker" }, **4**: **Self assessed general health**{"Bad","Fair","Good","Very bad","Very good" }, **7**:**Home Care Client Group**{"Dementia", "HIV, Aids, Alcohol Or Drugs", "Learning Disability" , "Mental Health Problems", "Other Client Groups", "Physical Disabilities"}, **8**:**Care Home Sector / Type Of Tenure** {"All", "Owned MortgageLoan", "Owned Outright", "Rented" }, **9**:**Household Type**{"Adults", "All", "Pensioners", "With Children" }, **11**:**Birth Weight**{"Live Singleton Births", "Low Weight Births"}, **12**:**Home Care Provider** {"Local Authority", "Private Or Voluntary"}, **17**:**Reason for Delay** {"All", "Code 9 Reasons (complex)", "Health and Social Care and Patient Carer Family Related Reasons", "Health and Social Care Reasons", "Patient Carer Family Related Reasons" }, **18**:**Alcohol Condition** {"Alcohol Related Brain Damage", "ALD - Fatty Liver", "All mental & behavioural disorders due to use of alcohol (M&B)","M&B - Psychotic and amnesic disorders", "Alcoholic Cardiomyopathy", "ALD - Hepatic Failure", "M&B - Acute Intoxication", "M&B - Withdrawal state", "Alcoholic Gastritis", "ALD – Unspecified", "M&B - Alcohol Dependence syndrome", "M&B - Withdrawal state with delirium","ALD - Alcoholic Hepatitis", "All alcohol conditions", "M&B - Harmful Use", "Toxic Effects of Alcohol"}, **19**:**Alcohol Related Hospital Activity** {"Value" }, **20**:**Type Of Hospital** { "All", "General Acute Hospital", "Psychiatric Hospital"}, **23**:**Age 1**{"16-34" ,"35-64", "All","16-64", "65 years and over "}, **25**:**Number and rate per 1,000 population of home care clients by disability condition** {"Value"}, **26**:**Main Client Group In Care Home** {"All Adults", "Adults with Learning Disabilities", "Adults with Mental Health Problems", "Adults with Physical Disabilities" , "All Adults", "Older People Aged 65 and Older", "Other Groups" }, **27**:**Care Home Sector 2** {"All Sectors", "Local Authority and NHS Sector", "Private Sector ", "Voluntary Sector" }, **28**:**Weight Category** {"Clinical - Healthy Weight","Clinical - Obese","Clinical - Obese & Severely Obese","Clinical - Overweight","Clinical - Overweight, Obese & Severely Obese","Clinical - Severely Obese ","Clinical - Underweight" }

Attributes names are in boldfaced letters while the levels are between curly brackets.

TABLE III
OBSERVED COUNTS AND TRIPLETS DENOTED AS (N)(x=NAME)($y.z$)

(21) AlcoholStats.(18.1,18.2,18.9, 18.10, 18.11, 18.12, 18.13, 18.14, 18.15, 18.16, 18.17, 18.18, 18.19, 18.20,18.3, 18.4, 18.5, 18.6, 18.7, 18.8) , **(4) AlcoholYoung**.(1.2,1.3,1.1, 21.1), **1 IntensiveHC**.Value , **(10) SingleRooms**.(8.1,8.2,8.3,8.4,26.1,26.2,26.3,26.4,26.5,26.6), **(1) GPBoard**.Value, **(9) MentalDisability**.(23.1,23.2,23.3,2.2,2.3,9.1,9.2,9.3,9.4), **(2) Lowbirthweight**.(11.1,11.2), **(1) GPWorkforce**.Value, **(4) HCDisability**.(23.1,23.2,23.3), **(1) HCServices**.Value **(3) HCArrangements**.(14,15,16).Value, **(7) HCProvision**.(27.1,27.2,27.3,27.4,26.1,26.2,26.3), **(7) Ensuite**.(8.1,8.2,8.3,8.4,26.1,26.2,26.3), **(4), DrugDischarges**.(1.1,1.2,1.3,17).Value, **(3) BMIDistribution**.(2.1,2.2,2.3), **(3) SmokingYoung**.(1.1,1.2,1.3), **(10) DDMonth**.(16.1,16.2,16.3,16.4,16.5,17.1,17.2,17.3,17.4,17.5), **(1) EmergencyAdmissions**.Value, **(3) BMIChildren**.(1.1,1.2,1.3), **(9) HCClientsGroup**.(7.1,7.2,7.3,7.4,7.5,7.6,2.1,2.2,2.3) , **(5) Occupancy**.(8.1,8.2,8.3,8.4,26.1)

The (N)'s are the counts observed per HSc triplet, the (name)'s or (x's) are the HSc's short names, the (y)'s are their attributes and the (z)'s are their levels using the HSc's naming convention in Tables II and I. Each triplet is a separate TS that spans over the 39 months or a specific dates group.

CC can work well for low sampled data (a few years) while χ^2 cannot be used well in this case (very few bins or just one). CC cannot work well for very variant data (very distant values) whereas χ^2 can accommodate these cases as different bins as is the case here. The data (services attendances) are actually transformed into their extended histograms and their definition domain (years that are attended) are re-scaled to include any extremes (outliers) of empty years.

V. PREDICTING THE CLASS

The CART, the RF as well and the LGR methods were used to predict the classes of all factors. This enables us to see what other H&Sc factors are usually involved when a target factor's class is studied. Class prediction links two or more H&Sc factors that help to predict it while hypothesis testing links two of them as likely to link or not and it can be extended to more H&Sc factors using the association matrix like the one shown in Fig. 1. A similar ML problem is addressed in works like in [12] as *'trend forecasting '*. The trend in the referred work is the 1^{st}linear coefficient that is the slope of the demand function and shows the primary predictorfor the target that mainly defines its trend.

For all ML classifiers and H&Sc factors the AUC and ACC metrics were used and the ground truth data were the cut attendances values with respect to the long-term average as can be seen in Figs. 5a-5c. The H&Sc factors that are associated with high confidence or high AUC and ACC to the target H&Sc factor to predict, can be members of the same group. We can train the 3 ML classifiers using a random set of years from the predictors and then test the classifiers using the same predictors with the test years. Or, we can use for training a sample set of H&Sc factors for all years and use the rest of the H&Sc factors for testing. Indicative evidence of ML or statistics-based relationshipsfound are presented in Table IV. Then the training H&Sc factors that give a high AUC or ACC and the predicted H&Sc factor form a lineargroup of services that one can offer as a plan to the public at less cost. For example the H&Sc factor *'LivingA . Gender . Male '* is well predicted by the H&Sc factors , *'HCClients. Main client group . in Care Home . Adults with Mental Health Problems '* and *'Lowbirthweight . Birth Weight . Live Singleton Births '* reaching among the highest AUC and ACC across the tests shown ($RF(AUC) = 0.949, LGR(AUC) = 0.923, CART(AUC) = 0.923, RF(ACC) = 0.926, LGR(AUC) = 0.923, CART(AUC) = 0.885$). The same combination of H&Sc factors works well using CART and LGR. The same holds for the combination *'AlcoholStats . Self assessed general health . Very good '* as a target with *'GPWorkforce . Gender. All '* and *'LivingA. Gender. Female '* as its predictors and the results are ($RF(AUC) = 1, LGR(AUC) = 1, CART(AUC) = 0, RF(ACC) = 0.912, LGR(AUC) = 1, CART(AUC) = 1$). These findings can help policy designers to see that the services offered to disabled patients can be a function of the type of house needed and of how this was obtained (for example by Loan/Mortgage). Housing quality is a typical indicator of good health. In the present work, both are found related and this is also observed in [13]. In the present work the quality of community housing is found directly related to the type of housing offered to disabled people or to people with social needs as can be seen by the connection of housing-related attributes in this segment of the population (for example community housing, care home housing, etc.). It is interesting to note that when a service is offered to disabled males then the connection to *'AlcoholStats . Self assessed general health . Very good '* is more evident (higher AUC's/ACC's). Another point is that singleton births are connected to mental health problems. It is sometimes challenging to understand how some relationships hold. Such an example is how *'AlcoholStats . Self.assessed general health.Very good '*, *'GPWorkforce . Gender. All '* and *'LivingA. Gender. Female '* can be associated. This is because it is not easy (obvious) to relate those admitted to a hospital after having consumed alcohol and said their health was very good to GP's of all genders and to those patients that receive care at home and are females. These are often called spurious correlations and sampled examples of them are listed in [14]. In [15] an example is discussed where COVID-19 death rates are related using ML (LR) to factors as diverse as socio-economic, county-level health variables, ways of

commuting, and climate and pollution patterns. Moreover, it can be expected, as in [16], that the patients who are most deprived (*'AlcoholY . SIMD quintiles . 1[=most deprived] '*) and are admitted can be related to adult patients that give low birth weights and are hosted in households or that Adults as in *'Mental. Household Type . Adults '* are related to those GPs offering services to households that are rented as in the H&Sc factor *'HeadcountGP. Type of Tenure . Rented '*. This is not observed, though, largely. Social deprivation of any kind (for example limited living space) may indicate low income level that may, in turn, be connected to low birth weights. Such a study that links social factors to low birth weight is reported in [17] along with other factors such as low iron intake and low mother's weight. In the present work the last reported relationship has average confidences (AUC and ACC) in the interval ([0.6,0.85]) that are low compared to other social factors relationships that are on the upper 90% range. The complex way by which the HC workforce can be affected or be predicted, using ML, by external factors (i.e. predictors) that are outside the HC setting (such as IT innovations) is discussed in [18]. For example the workforce and IT skills that are used and needed, nowadays, in HC operations (example keeping, updating managing EHRs) are not related to the rest of the common factors that are recorded during medical or social interventions. The data one had at hand in the present study are mostly concerned with factors that are part of that HC provision process and are not related to the IT skills needed to run an HC business.

Among the other ML methods used CART was used as a classification and regression method that learns. It is used both as a classifier and a predictor. The CART models can be trained to produce either a set of probabilities for a single or more input patterns and then it can predict their values. CART can be used as a binary classifier and a decision-making tool. In this work, CART was used as a method to classify the range of the H&Sc factors. The use of CART as a predictor of exact values had limited success with our data. This weakness was not observed, though, with the other classifiers. RF is an improvement to classification using CART but they are not as good as regression factors. An example is when fitting curves or in predicting continuous variables as discussed in [19]. For both CART and RF that are tree-based models, the benefit is that the decisions are split that allows for direct interpretation as reported in [20]. The ML methods were able to track with variable success the zero-padded data that are a common irregularity in the HC data analysis [21]. On the other hand, it is true that ML methods are biased as the result of not being able to guarantee that all cases or data classes are evenly represented in the training set as discussed in [22]. In our case, the zero-padding offered an additional potential bias that are the many zeros that have offered higher success rates. RF combines decision trees and converged less easily than CART models did in the experiments carried out in this work. The LGR was used to predict the probability that a H&Sc factor will next time (year) fall above or below its long-term mean. The formula for LGR, in this case, is given below:

$$P(HSCF_i \in C_k) = \frac{1}{1 + e^{-\sum_{j=1, j\neq i}^{j=m} a_{i,j,k} * HSCF_j}} \quad (3)$$

where $k \in [1, 2]$ are the IDs of the above- or below-average classes. The $a_{i,j,k}$s are called LGR coefficients and are learned during a training phase. The learned probability can then be *'thresholded '*(be cut below some threshold). Then the H&Sc factor is attributed or not to a group of other H&Sc factors that are used as its predictors. These are ML methods that are trained using subsets of the 110 dates sets and in some cases training years were used to predict the attendances for the rest of the years (test years).

VI. Social Factors Interaction vs. Colinearities

The H&Sc factors were studied so far as independent variables that can shape the target attendance provided that the independent H&Sc factors taken into account are not mutually linear. It is now assumed that a change of one independent factor's attendance does not only affect the target attendance but also the other independent factors ' attendances . Then the examined independent factor is said to be interacting with other independent factors' and it is called statistical interaction or quadratic interaction. It describes the relationship between two H&Sc factors $HSCF_i$ and $HSCF_j$ when they predict a target H&Sc factor $HSCF_k$ where $i \neq j \neq k$. The simplest such interaction / relationshipis given by:

$$HSCF_k = a_{i,j} * HSCF_i * HSCF_j + a_0 \quad (4)$$

Such an interaction between social H&Sc factors is also discussed in [23] where the target variable is the psychosis subgroup (taxonomy) that is dependent on H&Sc factors such as psychosis symptoms and depression symptoms as well as on the quality of life factors. A case of interacting clinical factors are studied in [24] where the patients are admitted to the ICU with prior hemodynamic compromise (example: IAC (indwelling arterial catheters)) that may be dealt with vasopressor support or concomitant sepsis as alternative reasons for IAC. The ICU admissions may be done for reasons that combine prior treatments as interacting interventions when the patients receive a single compromise or both. In the case of our H&Sc data the interaction of social factors was studied either by using the $\chi^2$and Student (t-test) tests or by using multiple single effect linear coefficients. An interesting point in the last referred work is that one can limit the interaction term by only considering subjects that have been explicitly treated with a single hemodynamic compromise. These patients can be identified using causal inference that relates the cause that is the prior treatment with the clinical outcome that is the ICU admission reason. Or, one can use a probabilistic method called *Inverse Probability Treatment Weighting* (IPTW) that links the clinical outcomes to the probability of being treated in some of the available ways. This is the propensity score that relates a clinical outcome to an independent factor which is its possible cause. In the present work, the propensity scores are more generically used in the H&Sc context and are identified by the outcomes

TABLE IV
REPRESENTATIVE RESULTS FOR CLASS PREDICTION USING RF, CART AND LGR, P-VALUES FOR χ^2 AND STUDENT (T-TEST) TESTS ACROSS ALL DATES GROUPS AND 39 YEARS

Linear groups of HSC's	kNN(AUC)	kNN(ACC)	RF(AUC)	RF(ACC)	LGR(AUC)	LGR(ACC)	CART(AUC)	CART(ACC)	MIE	CC	χ^2 (p-value)	[7] Stud.t.p.v.	Groups
HCDisability.26.3[1] , GPBoard.8.2, GPBoard.9.3	$0.7^{0.4}$	$0.8^{0.6}$	0.82	0.69	0.76	0.69	0.82	0.682	-0.6, -0.59	0.9, 0.9	0.0049, [6]**0.050**, 0.041	0.028, 0.0006, 0.0006	3,3,3
AlcoholYoung.2.3[1], HCDisability.26.3, MentalDisability.11.1	$0.7^{0.4}$	$0.5^{0.75}$	0.95	0.93	0.92	0.92	0.92	0.89	0.10, -0.92	0.25, -0.1	0.0005,**0.051**, 0.041	0.0007, 0.0006, 0.0004	1,3,3
$AlcoholStats.18.3_5$, Lowbirthweight. 2 . 2 , GPBoard. 8 . 3	$0.5^{0.8}$	$0.6^{0.8}$	0.82	0.78	0	0.66	0.82	0.68	-0.22, -0.26	0.35, 0.48	0.0035, **0.050**, **0.0005**, **0.0005**	0.0006, 0.03, 0.0004, 0.0	2, 4
AlcoholStats . 18 . 5^5, GPWorkforce. 2 . 1, HCDisability . Living Arrangements . 2 . 2	$0.75^{0.5}$	$0.8^{0.6}$	1	1	0	0.91	1	1	0.80, 0.19, 0.1	0.25, -0.1, 0.25	0.0005, **0.0505**, 0.041	0.0, 0.0006, 0.0, 0.0007	2,4,5
$HCDisability.27.4_1$, Lowbirthweight . 2 . 3, GPBoard . 9 . 1	$1^{0.7}$	$1^{0.7}$	0.82	0.78	0.82	0.78	0.82	0.68	-0.44, -0.44	0.48, 0.48	0.0005, 0.05, 0.041	0.0, 0.0015, 0.0	3,4
AlcoholYoung . 1 . 1[1], Lowbirthweight . 2 . 3, GPBoard . 8 . 4	$0.66^{0.6}$	$0.8^{0.62}$	$0.91^{0.7}$	$0.90^{0.2}$	$0.3^{0.6}$	0.66	$0.2^{0.82}$	$0.8^{0.68}$	-0.61, 1.92	0.99, 0.51	0.0005, **0.050**, 0.041	0.0, 0.0, 0.0006	5
GPWorkforce . 6 . 6^2, HCDisability . Living Arrangements . 2 . 1, GPBoard . 9 . 3	$0.8^{0.6}$	$0.8^{0.5}$	$0.92^{0.1}$	$0.9^{0.1}$	$0.97^{0.4}$	$0.97^{0.4}$	$0.89^{0.1}$	$0.8^{0.88}_{0^{0.4}}$,	-0.609, 0.188	0.51, 0.54	**0.050**, 0.0004, 0.041	0.0, 0.0004, 0.0004	5

The DGs(as indexed) next to the 1^{st}factor (1^{st}column) are not always the same as the actual ones (in last column) that are the original DGsof the services as in table I because they refer to the specific years a valid relationshipwas found. [1]2004-2010,[2]2010-2016,[3]2012-2019,[4]2014-2019, [5]1981-2019, x^y means best AUC/ACC x was obtained for learning ratio y when both $x <= 1$ and $y <= 1$, [6]boldfaced numbers under the p-value column indicate possible dependence among the $1^{st}$and the H&Sc factor with that order [7] Student (t-test) p-value: if($p - value <= 0.05$) then > 0 else $= 0$

TABLE V
INDICATIVE LIKELY CO-ATTENDED (LOW MIE) H&SC FACTORS

H&Sc factors names	Dates Group
HCDisability . 27 . 3 , HCDisability. 27 . 1	2004-2016
HCDisability . 27 . 2 , HCDisability. 26 . 2	2010-2016
HCDisability . 27 . 1 , HCDisability . 27 . 3	2012-2016
HCClientsGroup . 27 . 3 , HCClientsGroup . 27 . 4	2014-2016
HCArrangements. 2 . 2 , HCArrangements. 2 . 1	2014-2016
HCDisability . 26 . 5 , HCClientsGroup . 26 . 4	2014-2016

TABLE VI
MOST CORRELATED HSC FACTORS ACROSS DATES GROUPS

Representative H&Sc factors names	Dates Group
Lowbirthweight.11.1 , DrugDischarges.1.3, AlcoholStats.18. Value	2004-2016
Occupancy.8.4, SmokingYoung.1.3	2010-2016
BMIDistribution . 2. 2, SmokingYoung . 5. 1	2012-2016
BMIChildren . 1 .1, SmokingYoung . 1 . 1	2014-2016
GPWorkforice . 9 . 3 , MentalDisability SSCQ . 2 . 2, GPWorkforice . Value , HCClientsGroup . 7 . 6, ...	1981-2019
HCDisability . 1 . 3, Lowbirthweight . 10 . 3, etc	1981-2019
HCDisability. 26 . 4, HCDisability . 8 . 2, etc.	1981-2019
Lowbirthweight . 9 . 3 , IntensiveHC . Value, Lowbirthweight. 21 . 16-34, AlcoholStats . 18 . 13, AlcoholYoung . 21 . 2, etc	1981-2019

TABLE VII
INDICATIVE LIKELY CO-ATTENDED (HIGH χ^2 P-VALUE(>0.05)) ZERO-PADDEDH&SC FACTORS DEFINE SERVICES OR PATIENTS GROUPS

H&Sc factors combinations $HSCF_1$,...,$HSCF_k$,...	$p - value_k$ ($k >= 2$)
Primary 1 BMI Distribution . 8 . 1 , Primary 1 BMI.Distribution . 26 . 2	0.05
[1], Primary 1 Children BMI Epidemiological . 28 . 4	0.0555,0.052
[1], Number General Practices Registered Patients . 9 . 2	0.0525
Primary 1 Children BMI Epidemiological . 28 . 5, Primary 1 BMI Distribution . 26 . 1	0.0535,0.0670
Number General Practices Registered Patients . 1 . 3, Smoking behaviour and self rated health SALSUS. 28 . 1	0.0560
Primary 1 Children Body Mass Index Epidemiological . 28 . 3, Number General Practices Registered Patients. 23. 1	0.0510
Primary 1 BMI Distribution . 27 . 3, Primary 1 BMI Distribution. 26 . 7	0.0575

[1] Primary 1 BMI Distribution . Care Home Sector . All sectors

TABLE VIII
INDICATIVE COMPARISON OF (A) STATISTICALLY INTERACTING ($HSFC_i = a_0 + a_{j,k} * HSFC_j * HSFC_k$) AND (B) INDEPENDENT ($HSFC_i = a_0 + a_j * HSFC_j + a_k * HSFC_j HSFC_k$) HSC FACTORS WITH ASSOCIATED PROBABILITIES AND R^2ERRORS

$HSCF_i, HSCF_j, HSCF_k$,...	$pa_{j,k}$	$R^2(a)$	pa_j	pa_k	$R^2(b)$
BMIDistribution.2.1, HCServices.Value, HCProvision.26.2	6e-12	8e-01	0.008	0.97	0.231
AlcoholYoung.21.1, SmokingYoung.1.3, HCServices.Value	2e-05	0.99	0.230	2e-16	0.99
HCProvision.27.3, En-Suite.8.3, HCProvision.27.1	0.0024	0.996	3.9e-5	2.3e-15	0.995
SmokingYoung.1.3, BMIChildren.1.3, En-Suite.8.2	<2e-16	0.993	<2e-16	0.267	0.919
BMIChildren.1.3, SmokingYoung.1.2, SmokingYoung.1.3	4.2e-09	0.999	<2e-16	0.0024	0.999
AlcoholYoung.1.1, SmokingYoung.1.3, En-Suite.8.1	0.005	0.391	0.0067	0.921	0.231
HCClientsGroup.7.3, HCProvision.26.3, GPWorkforce.Value	0.003	0.302	0.248	0.360	0.114

The factors naming convention here (i.e. acronym.X.Y) uses the H&Sc acronyms from table I while the attributes IDs and their levels IDS follow the ordering in the tables II and III

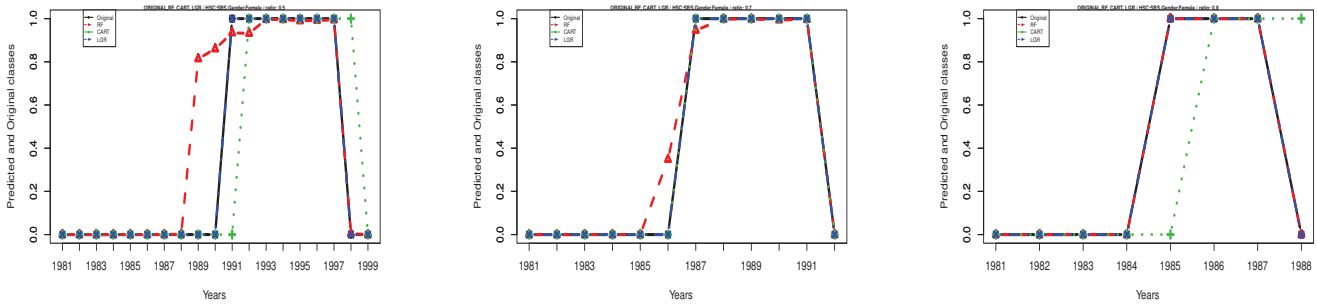
of the LGRor other ML methods that explain why a target attendance reaches a level above its long-term average. Hence the LGR, RF and CART methods may also describe factors interactions. In the case of LGR, (3) may assume a form like $P(HSCF_i \in C_k) \approx exp(-a_{i,k,l,n} * HSCF_l * HSCF_n\})$

Such indicative interactions of factors are shown in Table VIII where the first factor is the target and then the rest are the predictors that interact both one with another and jointly with the target. The table is limited to show interactions of pairs. One can observe that for several of them two from the 3 terms used are from the same social groups (for example *'HCProvision '* or *'SmokingYoung '*). Or, they concern the same population segment (for example elderly people: *'HCClientsGroup '* and *'HCProvision '*). One can also say that the prediction errors (for example the R^2) considering the interaction and the non-interaction between the factors are similar but the confidences in the interaction coefficients are stronger whereas the non-interaction probabilities are lower. That means that the predicting factors are not entirely independent. The interaction of prediction terms in HC spending can explain the risk associated with the inability to accurately predict claims as discussed in [25].

It can be stated that either of the two interacting factors is co-founding one with respect to the target. The role of the confounding factors in this work is rather loose since the data were randomly selected and are not derived from organized clinical trials databases [24] nor are focused on social studies with specific questions and audiences. Such a setting would make it more necessary to study how specific factors relate in different ways as well as the way they affect the target. This implies co-founding, i.e., one factor can have an unknown impact both to the result and to another factor and then determine how the 2^{nd}factor predicts the target.

VII. CONCLUSION

The paper discussed how can we group H&Sc services by classification and regression and a comparative study was presented including ML and two major statistical hypotheses tests, χ^2 and Student (t-test). It was advocated that the prediction that one member-service of the group will be



a b c

Fig. 5 (a)CART, (b)RF and (c)LGRclassification results for HSCF *'Smoking behaviour and self rated health SALSUS.Gender.Female '*. The decision vectors are shown in black solid lines the RF classificationis in red line the LGRis in blue line and the CART is in green. The CART under-performs with respect to Random Forests especiallyfor low ratios while for higher ratios and better training CART seems to be picking up well the H&Sc factor . LGRshows very good performance

attended above or below its long-term average can be based on the attendances of other member-services of the same class. The groupings can change in kind, confidence, and over time. The long-term average was the cut-value used but one can also use attendances ups and downs over years as a criterion for the classification decisions. Different metrics were used to find cohorts or to predict these probabilities using various H&Sc factors sub-sets. Groupings-based relationships were studied using LGR, CART and RF that refined results obtained from tools such as CM, CC, and k-NN. The hypothesis tests verified some of the ML supported relationships found taking the p-value of the hypothesis and the information exchanged between the data series. The initial groupings found were similarity-based and relied on CC or CM and then the results were further explored using ML methods. CART did not work as well as a classifier compared to RF and LGR in predicting the probability sought. For a better analysis, specific dates intervals were used that favoured shortly attended services and some regression types that did not have more than 1 or 2 predictors and so did the entire period of 39 revealing nevertheless the changing nature of those relationships. The dates groups were less interesting due to the limited data obtained with common years to check complex H&Sc factors relationshipsthat were better studied in the long-term. At the same time, both time frames helped to better understand data in the long-term as well as in the short term. The admissions due to Alcohol were the most prevalent H&Sc factor. Using the dates groups as a framework the detailed analysis was made possible and specific relationshipswere found while the 39 years setting helped to understand how HC services change in the long run that could not be observed otherwise. Especially the LGR method provided a probabilistic framework that proved that services are uncertain and may depend on many factors and especially on the time the data were recorded. Generally speaking, some H&Sc factors were found to be widely attended such as the Emergency Department services that were highly cross-correlated to other less attended H&Sc factors and over many dates groups. The paper did not focus, though, on specific H&Sc factors since many relationships proved to change over time. The intention was less to provide sociological insights as was to provide a computational tool that may help.

ACKNOWLEDGMENT

The author is grateful to PHS and to NHSS Scotland and especially to Mr. Martin McKenna, Mr. Andrew Mooney, Dr. Lee and Mr. Paul Leak (Scottish Government) for accommodating in their technical meetings, for sharing ideas, bibliographies that helped and guided this work. The author is also thankful to Dr. James Bown, Dr. Euan Dempster and Dr. Ann Savage from Abertay University Dundee for comments in early phases of this work. The author during the works of this paper was funded by an Abertay University Dundee.

REFERENCES

[1] Scottish Government, Statistics Service Health and Social Care Data:Growing up in Scotland: health inequalities in the early years. statistics.gov.scot. https://www.gov.scot/publications/growing-up-scotland-health-inequalities-early-years/pages/5/

[2] Nnoaham K E, Cann K F Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population?. BMC Public Health. 20, 798 (2020). https://doi.org/10.1186/s12889-020-08930-z

[3] Benjamin Seligman, ShripadTuljapurkar, DavidRehkopf Machine learning approaches to the social determinants of health in the health and retirement study . SSM - Population Health. volume 4, April 2018. Pages 95-99. https://doi.org/10.1016/j.ssmph.2017.11.008

[4] Ian Litchfield Can process mining automatically describe care pathways of patients with long- term conditions in UK primary care? A study protocol. BMJ Open. 2018. https://bmjopen.bmj.com/content/8/12/e019947

[5] Bose, Johnson, Alistair, Moskowitz, Ari Celi, Leo (2016). Raffa, Jesse. (2018). Impact of Intensive Care Unit Discharge Delays on Patient Outcomes: A Retrospective Cohort Study. Journal of Intensive Care Medicine. 34. 088506661880027. 10.1177/0885066618800276

[6] Rahmandad, H.,Oliva R.,&Osgood, N. D. (Eds). (2015a). Chapter 1: Parameter estimation through maximum likelihood and bootstrapping methods. In: Analytical methods for dynamic modelers (pp. 3–38) . MIT Press.

[7] Liying Fang, Han Zhao et al. Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data. Biomedical Signal Processing and ControL. 21 (2015) 82–89. https://core.ac.uk/download/pdf/82644081.pdf

[8] van der Hoef H, Warrens M J Understanding information theoretic measures for comparing clusterings.. Behaviormetrika. 46 353–370 (2019). https://doi.org/10.1007/s41237-018-0075-7

[9] Claudio Heinrich. On the number of bins in a rank histogram(2020). 2, https://arxiv.org/pdf/2005.09018.pdf
[10] Myers, P.D., Ng, K., Severson, K. et al.. Identifying unreliable predictions in clinical risk models. npj Digit. Med. 3, 8 (2020). https://doi.org/10.1038/s41746-019-0209-7.
[11] Lokhandwala S., Rush B. (2016) Objectives of the Secondary Analysis of Electronic Health Record Data. In: Secondary Analysis of Electronic Health Records. Springer, Cham. $https://doi.org/10.1007/978-3-319-43742-2\_1$
[12] Xu S., Chan H K , Ch'ng, E. et al. A comparison of forecasting methods for medical device demand using trend-based clustering scheme (2020). J. of Data, Inf. and Manag. 2, 85–94 (2020). https://doi.org/10.1007/s42488-020-00026-y
[13] Health 2020: Social protection,housing and health - September 2016. World Health Organization (WHO), $https://www.euro.who.int/\_data/assets/pdf\_file/0005/324635/Health-2020-Social-protection,-housing-and-health-en.pdf$
[14] Spurious-correlations examples . http://tylervigen.com/spurious-correlations
[15] Christopher R. Knittel, Bora Ozaltun. What does and does not correlate with COVID-19 death rates. medRxiv 2020.06.09.20126805 . https://doi.org/10.1101/2020.06.09.20126805
[16] Scottish Government, Statistics Service Health and Social Care Data:Growing up in Scotland: health inequalities in the early years. statistics.gov.scot. https://www.gov.scot/publications/growing-up-scotland-health-inequalities-early-years/pages/5/
[17] K C, Anil, Basel, P. L., & Singh, S. (2020). Low birth weight and its associated risk factors: Health facility-based case-control study. PloS one, 15(6), e0234907. https://doi.org/10.1371/journal.pone.0234907
[18] ARI BRONSOLER, JOSEPH DOYLE, JOHN VAN REENEN. The Impact of New Technology on the Healthcare Workforce. MIT work of the future, Research Briefs - October 2020. https://workofthefuture.mit.edu/research-post/the-impact-of-new-technology-on-the-healthcare-workforce/
[19] Md. Zahangir Alam A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked(2019). volume 15, 2019, 100180. https://doi.org/10.1016/j.imu.2019.100180
[20] Breiman, L. (2017). Classification and regression trees. Routledge.
[21] Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data:A systematic review. Journal of the American Medical Informatics Association, 25(10), 1419–1428.
[22] Chen, P. C., Liu, Y., & Peng, L. (2019). How to develop machine learning models for healthcare. Nature Materials, 18(5), 410.
[23] Dwyer DB, Kalman JL, Budde M, et al.. An Investigation of Psychosis Subgroups With Prognostic Validation and Exploration of Genetic Underpinnings: The PsyCourse Study . JAMA Psychiatry. 2020;77(5):523–533. doi:10.1001/jamapsychiatry.2019.4910
[24] Hsu DJ, Feng M, Kothari R, Zhou H, Chen KP, Celi LA. The Association Between Indwelling Arterial Catheters and Mortality in Hemodynamically Stable Patients With Respiratory Failure: A Propensity Score Analysis. Chest. 2015 Dec;148(6):1470-1476. doi: 10.1378/chest.15-0516. PMID: 26270005; PMCID: PMC4665738.
[25] Irvin, J.A., Kondrich, A.A., Ko, M. et al. Incorporating machine learning and social determinants of health indicators into prospective risk adjustment for health plan payments. BMC Public Health 20, 608 (2020). https://doi.org/10.1186/s12889-020-08735-0