World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:18, No:9, 2024

Designing Social Care Plans Considering Cause-Effect Relationships: A Study in Scotland

Sotirios N. Raptis

Abstract—The paper links social needs to social classes by the creation of cohorts of public services matched as causes to other ones as effects using cause-effect (CE) models. It then compares these associations using CE and typical regression methods (LR, ARMA). The paper discusses such public service groupings offered in Scotland in the long term to estimate the risk of multiple causes or effects that can ultimately reduce the healthcare cost by linking the next services to the likely causes of them. The same generic goal can be achieved using LR or ARMA and differences are discussed. The work uses Health and Social Care (H&Sc) public services data from 11 service packs offered by Public Health Services (PHS) Scotland that boil down to 110 single-attribute year series, called 'factors'. The study took place at Macmillan Cancer Support, UK and Abertay University, Dundee, from 2020 to 2023. The paper discusses CE relationships as a main method and compares sample findings with Linear Regression (LR), ARMA, to see how the services are linked. Relationships found were between smoking-related healthcare provision, mental-health-related services, and epidemiological weight in Primary-1-Education Body-Mass-Index (BMI) in children as CE models. Insurance companies and public policymakers can pack CE-linked services in plans such as those for the elderly, low-income people, in the long term. The linkage of services was confirmed allowing more accurate resource planning

Keywords—Probability, regression, cause-effect cohorts, data frames, services, prediction.

I. Introduction

THE concerns about the system of publicly funded social care in England and in Scotland can be alleviated with the use of combined modern methods. Analysis tools have become recently available to face the growing demand by connecting services using CE service models and, also to examine the use of typical prediction methods. In these models, we have one or more services that can be considered as the outcomes (effects) or as targets of one or more other services that are modeled as their "causes" or as their predictors. The motivation of the proposed approach is to define service relationships by using the flexibility of the CE models and the exact formulation of regression models. CE models are quite generic and might help to draw conclusions from these service relationships using causal inference (CI) while LR and ARMA can be part of certain CE models and offer a precise mathematical context.

The CE models and the regression methods can be trained by using the available data. The CE model parameters can flexibly link different kinds of services to use as pre-cursors to the next services and the regression models link a clear target to specific predictors. The implementation differences and the similarities are observed and how they can help with policymaking. The CE models used in this work cannot hold across the entire year-span of the data, that is, the 39 years. The limited time over which one can link them can be the period of time in which one can more safely link a service's demand to other services' demands. Still, there is an associated confidence level attached to it. That means this probability can help to find out how well they can be linked over this defined period. A further motivation for this work is that one can extent services CE linking in policymaking. The CE models can help to determine the effects of the policies applied on some target variables of interest which are the measurable outcomes of the applied policies. Such can be different living quality factors such as how many people take a service after a policy is applied. Policymaking can be assisted by using such analysis methods. One can easily extend a methodology that is typically applied in a clinical setting to the social setting. Policymaking can also be seen as a data process considering how many people took one or more public services before and after some policy was applied. Then, one can measure how this changed after a political decision or new law is applied. In a clinical setting, there are different indicators that can flag the status (administrative, clinical, other therapy-related) of a patient. In clinical projects (interventions) one can broadly think of the risk of deterioration of one's health after a medicine is taken.

The paper is organized as follows: In Section I the motivation for pursuing this work is explained as a way to deal with cost reduction and for a better organization of the resources using CE models. Then, in Section I-A the relevant literature on CE models is given as well as an introduction on regression, prediction and CE models as service association methods. The time lag that applies to the current data (between nonzero records) is explained in Section I-B while the regression that needs time overlap is contrasted to the CE models that do not in that respect. The nature of the data processed is better explained in section II while the main analysis is given including the CE models and the rationale behind them as well as of the connection to LR and ARMA methods in Section III. Next, the data are better explained and the notion of the service pack, (H&Sc), is contrasted with the concept of the single data stream tracked, (TS). It is further explained how this facilitates detailed data linking using CE. Also, the common types of CE cases are given while this analysis is framed by the public services context in which the work is conducted and specific links are provided to put the present work in the CE's context. Thus, the application on PHS data (Public Health Scotland) is explained. The section also briefly presents the ARMA/AR prediction and LR models and discusses how these can work along with CE pairing while in

Dr. Raptis was during the works of the paper with the School of Design and Informatics, Abertay University, Dundee, Scotland (e-mail: sotiris.raptis.n@gmail.com).

the same section, the CE map and the specifics of pairing a cause with an effect through an intervention are given. Indicative comparisons and results are presented in Section IV and are accompanied by comparative co-plots or tabular forms. Emphasis is given to the link of CE, with LR and ARMA as ways to group service data (how they are provided) and customer data (to whom) in cohorts that facilitate the better organization of public services in the broader sense.

A. Related Works

The work in [1] discusses the concerns in regards to the cost of healthcare services in Scotland. These concerns can be faced by better organizing service provision using parameters (variables) that describe the services. The work in [2] discusses the role of hidden variables in making sense of observed clinical effects when the results are not clear.

In [3] an interesting point is made that the CE models can promote evidence-based decision-making. That is, link the evidence as measured (encoded) by cause variables to the effects that are the decision on actions to take as interventions or as options. Decision-making in CE studies may concern the best action to take considering causes (observed evidence) as well as ways to gauge the effects if the effects cannot be made more precise. An example is the quality of care, or, the quality of life that are multi-parametric effects as discussed in the study in [4]. In the last referred-to paper, the quality-of-life factors taken can be psychological, lifestyle, and social parameters. This is central in estimating causal impacts, and in comprehending unintended consequences. Indeed, as discussed, if one expands the CE analysis to a wider spectrum of co-variates and considers as many as possible co-founding factors then the risk of having adverse effects is limited; as the work in [5] assesses in drug safety research the risks from taking a drug can be computed using the appropriate CE models to point out a product's adverse reactions.

The topic of CE modeling is widely addressed in the literature. Typically, CE methods are mostly used in clinical treatment decision-making as in [6]. In such a clinical setting the challenge of linking many clinical causes (for example, the administration of drugs) to many effects (for example, the clinical outcomes) is discussed. The CE models may also be applied to policymaking as in [7] where CE relationships between policy actions and their outcomes are the topic. In this context, an intervention can be regarded as a data process. In the same way, publicly offered services in the health and social care (H&Sc) setting, can be regarded as social or public interventions in the sense that people who benefit from them can make their life better (change their living quality). This can be measured in a plethora of ways. A case in the EU (European Union) is analyzed in [8] by the use of the suitable (KPI)s (Key Performance Indicators). The referred-to work is based on pre-set (H&Sc) policy goals that were to be attained. The CE pairing in CE models can also mean (not always) that a change or not of the demand of the effect (target) services can be sensitive (influenced) by changes that occurred in the cause-services that preceded them. As discussed, an example of service pairing can be how the housing market can change when the rents are partially or fully subsidized. The tenants can be charged less for rent due to an allowance in the rent offered as a social policy. Hence, this social policy can have an impact on the housing market.

A discussion regarding the UK housing market is presented in [9]. The referred-to work presents the full spectrum of effects either as benefits or as disadvantages. Both are still effects. Another example of CE pairing can be how hospital admissions depend on alcoholism. Alcoholism makes people seek help and be admitted to hospitals. Then, by using CE models, one can predict this number using CE, and by that one can also foresee seasonal admission peaks for some reason. In some areas of research, the CE models can only be useful when there can be a specific time-lag between the causes and their effects as in climate models, as discussed in [10]. In the last referred-to work the CE models that are based on Granger causality are used as especially applicable for time series (TS) analysis. The Granger test uses ARMA (Autoregressive Moving Average) models to link causes and effects. The time relationship between causes and effects is discussed in [11] where the volatile nature of CE relations is advocated and its dependence on the dynamics of the studied system (a chemical in the 'the referred-to work). The physical or chemical processes are typical examples of CE models but in this work, these can be inferred using ML methods such as prediction. The model used in this work is introduced in [12]. We can refer to different areas where the CE modeling can be applied in order to provide more context on how the CE modeling can be used to plan. The work in [13] makes a clear definition of risks conceived in the context of CE analysis that are applied in classical construction engineering. The effects are not only good ones but can also be adverse effects that either do not help a project to progress or may as well cancel past progress made.

In [14] the aspect of CE that deals with sensitivity (dependence) of the effect on causing interventions is presented. This goes beyond the likelihood of linking a cause to its effect and examines the probability of the existence of such an association as well as the of the credibility of it. The CE model itself is determined by a set of parameters that usually depend on the learning data set from which the model emerged (derived). Then, the sensitivity would depend as well on such parameters.

The CE models can link public resources used (IT, man-hours, allowances, etc.) by the H&Sc system to better deliver H&Sc services and alleviate the servicing cost. This cost is predicted to become as high as £12 billion by 2030/31 at an average rate of 3.7 percent a year as analyzed in [1]. As the work in [15] discusses there is a risk in considering more causes to an observed effect. CE analysis can solve a dual problem. The first is to well understand what brings a high outcome cost. An example is to have no reasoned expenses that could have been predicted or avoided as adverse effects. The second is to find the best (and cheaper solution) for a given positive effect that can be the goal of the policy. An example can be the early discharge from a hospital or fewer resources to provide as a social benefit when we know from the CE analysis what can cause the demand for it.

The work in [16] gives the spectrum of methods that are used in CE. In the counterfactual theory of causation, one can predict the eventual effect without the service that causes it. The counterfactual effect can then be computed using LR (linear regression). With LR one can predict the effect if we remove from the list of N causes in (2) the one that we wish to compute the counter-effect for. That is, without having demand data for that service. The potential outcomes framework that is also presented in the referred-to work is the inverse methodology of the one followed here.

In [17] a more formal setting of the CE models is given that also helps to categorize the approach taken in this work. The three basic quantities involved in CE models are:

- The cause-variable, X, which here, is the public service which causes another effect-service, Z, to start having clients (that is, people who benefit from Z). Z's demand levels rise as a result of the cause-service precursor.
- A counter-effect variable that is, a service, X_1, which is the version of the original service X, as it would evolve if the intervention, Y did not occur. Hence, the CE model causes the theoretically expected X_1 variable to be transformed to a new new variable, the Z that is the one observed. A metric that can be applied to measure the counter-effect is to take the average difference $X_1 - Z$ over all kinds of (Y)s.
- An effect variable Z that is the observed variable that the CE models

It is common to apply CE models in the healthcare (HC) as discussed in [18]. It is there advocated that changes administered in the public vaccination dosage can cause effects to the outcome as an observed variable. Such can be the generic health. The effects can also be on the co-founders (causes) that are clinical or social factors that come into play.

As the work in [3] explains CI helps to spot connections between services that ML cannot capture.

1) Relation between LR and CE: An interesting direction on how LR and of CE models can be related is given in [19]. It is stated that the direction of the prediction defines what causes what else. The referred-to paper then assigns a confidence to each direction. If we invert the direction of inference then the predictor becomes the predicted variable and the target becomes the predictor. This again can be assigned a confidence. The paper says that the direction that is assigned the higher confidence is the true CE model. A similar view of this is discussed in [20] where the the inference is modeled as a chain of predictable variables where one can predict the next one till the final target is approached. . The idea discussed is that an economic outcome can be the chain prediction of a sequence of factors where each depend the one preceding it. One can thus understand that these two concepts, CE and LR, are well inter-related although not the same.

The major prediction methods are reported in [21] such as random forests (RF). In [22] the causal forest is used as a method to compute the heterogeneity (variety) in treatment effects that are the individual-level glucose-lowering responses in clinical trials, conducted on subjects. The RF provided insights as for which therapy likely caused the effect using decision trees. In [23] LR is used to predict the work-load in hand surgery operations in aging population or it is used to predict the clinical outcomes in [24]. The roles of the different clinical factors as service attributes are analyzed in [25], [26] and in [27]. This can be relevant to choosing the most likely cause or role of them given that a predicted outcome is observed.

A direction of how LR is used in CE is analyzed in [28] as a way to find co-variates (causes that one relates one to the other). In this case, one cause-service can be predicted (but not necessarily inferred as an effect) by the rest of the causes. The counterfactual data are ”virtual” data, that is, not observed data that are usually the mathematical (modeled) counterparts of the observed as if the cause-data (services or events) did not exist which caused the observed ones to occur.

The social or public effects can be the benefits of applying social interventions. These can be quite distant from the time the interventions (social policies) were applied.

An adverse effect may be part of an unknown relation between a cause-variable and an effect one. In [29] the role of volunteers is discussed in determining the gains (effect variables) from a range of interventions (cause-alternatives) so that policymakers can make informed decisions. These decisions can concern how the healthcare services (and therapies) are delivered or chosen out of large lists of candidates. The analogous to social care plans, that are examined here, is how social care services are taken up by beneficiaries.

The CE models can conceptually be linked to LR and ARMA models. In ARMA models we have a self-generating service and one can define the effect directly from the past samples (for example, before 2011) of the same service and not seek for causes in other data streams.

Linear prediction is discussed in [30] while the linear association of different service parameters is also the question in [21] and a review of linear methods in healthcare (HC) is presented in [31].

B. Time-Lag in Open NHSS Data

In [18] an example of the time-lag between the point an intervention is applied and the time (till the onset of the COVID) as an effect variable is discussed. The referred work elaborates on the concept of the “counterfactual” effect variable. That is, the Cox regression model for the time of survival is applied and two results are compared over the same time. One model for no vaccination and one for the fact of vaccination. The time granularity of our data at hand (39 years) does not allow for a depth of analysis. That is, we choose different time lags and the effects of that. One can see, though, that the CE relationships across the services do depend on the period (group of years) considered. Maybe one can measure how one service can cause different services (as its effects) so long as the effect services can be defined using temporal CE models. In [32] the averaged treatment effect (ATE) is discussed as a way to assess the change that a cause or an intervention can bring to an effect service. ATE implies a changing CE model in time or suggests that the CE model may not apply for the entire treatment period. In this work we would need to take the average response of the likely effects from year #1 (1981) to year #39 (2019). Due to the heavy zero padding a 53.8 %of the 110 services had at least three records from 1981 to 2019 (that is the most frequent year span met in the data). That means that the CE models linked services of quite varying duration. The average year span for those services that lasted more than 1 year was 10 years. That also means that the CE models likely linked as causes 10-year services as effects to 10-year services as their causes.

We assume that the public services data which are the actual cause service data (X) and not the (X_1)s that are not really observed. We cannot really observe the counter-effect variables (that is, the "unaffected" versions of the original services (X_1)). In this work, if we call ($D = X_1, ..., X_{N=110}$) the data, then, the ($\{Z_i, Y_i, X_i\}, i \in [1, 110]$)s, are the cause-intervention-effect triplets whereas $X_{1,i} \notin D$ are the counter-effect variables. The common duration for all is 39 years and has zero-padded data. The CE horizons (durations), (d_i)s, are defined separately for each of all the thus-defined possible CE triplets that can be defined in our data. These durations are those for which we can have a confident CE model or a duration for which we can compute confident Bayesian parameters as those defined in (2). That means we can link subsets of public services in the data, D when these subsets meet the requirements of the CE models and the horizon of the years allows to compute them. This learning process means to learn them during the time defined between $\{t_0, t_1\}$ as in (5). The time t_0, is the point from which one can apply learning. The CE models can have separate learning horizons that vary from 10 to 20 years out of the 39 years in total. We found that given the extended zero padding applied, one could define CE models after year #30 (or 1981+30=2011) and for an average horizon $\hat{d}_i, i \in [1, 110]$ of 5 or 6 years. During the application horizon, one could link, using CE models, public service demands as effects coming from one or more public services that mostly preceded them.

Most services were sparse (many missing records) and started having digital records after 1997 and many had quite sparse records before that. Hence, the year overlap was stronger after 1997 which means that CE relationships as in (2) can stand (be non-zero) only for common years. An illustration of the data at hand can be observed in Fig. 1.

Our data had 42 %single raising or single stopping services which means that almost half of the public services examined stopped or started when no other service was offered. That means that while a cause service was been offered to the public some people decided to take one or more effect services that were not taken before. The percentage of services that started while others were already offered is 58 %. One can assume that there are co-founders, that is, servicesthat can link separately two or more servicesas separate CE pairs but have no common times between their start and end points That is, when people ask for a public benefit, then a new service starts as the result of a problem that is recorded as a cause service. Then, this does not mean that the effect service will start being offered immediately after that. A time interval in years may pass before the new service is set up.

The missing years between start and end points can create time-lags. Relevant methods for dealing with missing data can be found in [33]. The data imputation using Markov models as in [34], [35] can replace missing data. It is though beyond the scope of the present work to examine whether artificially recovered data can affect CE links.

II. Materials

1) The Data Landscape: The present work examined 110 different service demand data as raw data. Then, these data were aligned in time, with zero padding, so that one can have all service demands for a specific year from 1981 to 2019. However, NHSS initially packed these service demands and posted them on-line as offered to different cohorts of people in Scotland. These cohorts were computer-organized as data-cubes. This is also obvious from looking at the related data cubes data on the NHSS web-site [36]. Part of the time-alignment process was to apply zero-padding for serviceswith missing records or that did not have records in one or more years. A list of the 11 main services (service packs) that were broken down into their attributes (conditions they were offered in) is shown in the table in Fig. 2. The acronyms used per attribute and level are in parentheses. The attributes are boldfaced and their levels are separated by commas inside the parentheses. Some attributes do not have levels and are only reported as "Value". A service pack can have as many TS as each of its components (that is, separate year series/TS per attribute/level). Fig. 1 shows the sums of the counts of the demands per H&Sc (pack). That is, in this figure one sums-up the demands across all attributes and levels of them for each of the (H&Sc)s. As it will be discussed later, the work found that health and social care servicescan be combined through CE models in a multitude of ways. This suggests that, with a variable degree of confidence, one would expect that very different causes (as services that precede other ones) may lead (or be followed) by a quite diversified spectrum of other health and social care services. The later ones can be taken/modeled as their effect ones using the appropriate CE models. In fact, the heavy zero padding before 1997 (Fig. 1), when most later health and social care plans were not offered as such, adds to this. That is, many contemporaneous health and social care plans started having records (or started being offered) only after some year-point. Hence, this biases the data generation process towards a process where all data were generated after some point and not before for some reason. A H&Sc service can be again a data generation process where one starts creating/keeping records for services. A question that can arise is why not take only non-zero TS or after defining as a start point a specific year (if not 1981) where most of the (H&Sc)s (thus TS) would have records. In fact, there are, even in 1981, very low populated (of very low demand, that is) services that cannot be shown in the collective diagrams here. Indeed, the later recordings of the demand were of 3 or more scales higher. For example, the TS "BMI Distribution in Primary1 Children - Client Group of in Care Home (Home sector/voluntary sector)" or (HSC#6, Service ID:79) had (176944) hits whereas TS "Home Care Services / Value" (HSC#11, Service ID:110) had (5). It is worth studying

Open Science Index, Computer and Systems Engineering Vol:18, No:9, 2024 publications.waset.org/10013811.pdf

CE pairs or CE relationships in such a long span since CE pairs do not have to be adjacent in time. That means that the result of an applied H&Sc policy can manifest itself in the short or in the longer term. It was found that CE links mostly health and social care services to other ones and that only quite a few among them were not effected of some other service. It was also true that the health and social care data examined emerged from a few basic ones, that is, services that were tailored to specific audiences such as: (1) low income people, (2) adults, (3) young people smoking, (4) young people suffering from alcoholism, (5) the elderly, (6) very young mothers, (7) people with mental health issues, (8) people who receive care from distance or at home, (9) newborns with very low weight, (10) people in need of community housing, etc. That means that health and social care beneficiaries were naturally grouped in CE relationships from the way the data were generated. An example of that is that female adults of some age were linked (using CE) to male adults of similar age who lived in subsidized housing. Hence, adults of the same age regardless of gender needed free or subsidized housing and their numbers were very comparable (CE linked). We had data sub-sets that belonged to such divisions and were tracked as separate data sets. That is, as population segments need special health and social care plans for them. The CE modeling proved that such cases of similarly behaving population segments can be studied in conjunction, that is, by using a CE model. One can then think that one such service is the cause and its counterpart (as for age band or gender) is the effect service so that one can predict the demand for the effect service when knowing the demand for the cause-service. Moreover, one can link services of quite different natures such as services offered to low-income people as causes of services related to subsidized housing as an effect social care service.

A. The Data Format

The data we had were public H&Sc data available on PHS' [37] website [38] posted by June, 2019. The data used were counts of patients and are represented in the data under the attribute called ‘Value‘.

The data came in various formats for dates or for other counts (heterogeneous), with missing years, numerical tags, or bands for ages. Also, flags were used denoting the presence or not of a demographic class or of age-bands. Also, categorical data (social or demographic groups) or long text descriptions for them instead of numbers were found. For example, ages were found both as ranges as in ‘...ages 65+ ‘ or as single numbers. The gender would have a numerical tag ‘1‘ for ‘Male‘ and ‘2‘ for ‘Female‘. Other records were counts of patients (without more specifications) or percentages. A breakdown of the indicative attributes and their levels per attribute is shown in Table I. The data contained up to 6 attributes (service settings) per service and each attribute has possible values called ‘levels‘. The services without settings had one attribute (‘Value‘). Some data take up to 20 levels (as in 'Hospital Admission Reasons').

Acronyms such as ‘S.A.Z.‘ are used to denote factor names that are part of service packs (the (H&Sc)s). The 1^{st} part, 'S', of the acronym is the ID of the H&Sc pack, then the acronym, 'A', of the attribute follows and then the ID of the level of the attribute follows as 'Z'. For example the service ‘Alcohol use among young people ‘ is ‘S1‘, and the attribute for age 'A' has levels 'Z's: $\{‘13‘, ‘15‘, ‘All‘\}$. Each level is attached to a single data stream (a TS) that is processed and tracked or modeled as a service. That is, it was an individual factor or setting. This example indicates the number of patients aged 13 or 15 or of any age ‘All‘ whose number is tracked over the 39 year span. The services are also referred to in this work by their short names using the table in Fig. 2

III. Research Method

A. The CE Model

The typical CE model is a CE matrix that links 110 services (as possible causes) to 110 other services (as possible targets). It is given in (1):

$$M_{CE} = \{P_{i,j}\}_{110,110}, \ni P_{i,j} \in [0,1] \qquad (1)$$

It is a probabilities matrix. In this matrix, each row represents one target service and all its columns contain the attached confidence that each of the 110 services (the target is included) can cause the target. The pairs with the same (ID)s (that is, along the diagonal) can indicate that one service can, by definition, cause itself with $P_{i,i} = 1, i \in [1, 110]$. This needs to be contrasted to the case where one service is only part (that is, not alone) of the group of cause-services, which is also possible, but with $P_{i,i} < 1, i \in [1, 110]$. This was also discussed previously in the case of ARMA modeling self-prediction. Fig. 3 shows indicative rows (the (ID)s on the left box) that are the effect services. On the right of each plot in each plot are the cause services. The common year span in the sub-figures of Fig. 3 (x-axis) is 39 years ($[1, ..., 39]$)

The CE model is defined for a cause service's time series (TS), $\vec{x}(t) = [x_1(t), ..., x_N(t)]^T$, and for the observed effect's TS at the time point (t) with the form:

$$y_i = \vec{x_i}^T * \vec{\beta} + \epsilon_i, \ni i \in [1, N] \qquad (2)$$

for $\vec{\beta} \in R^{1 \times k}$ and $\epsilon_i \approx N(0, \sigma^2)$ that are IID variables.

The elements of the CE matrix are computed from (2. A CE model typically links an effect service to one or more cause services. The terms “effect” and “cause”, (CE pair), may obtain different names depending on the application. In clinical trials it can be a “dose/response” pair or an “intervention/outcome” pair or an “exposure/outcome” pair. A CE model assigns confidence to any such pair and this confidence determines to what extent is the relationship valid. Fig. 3 illustrates such CE combinations for two different indicative thresholds, $P_1 = 0.1$ and $P_2 = 0.5$. The exact model for binary (CE)s is given in (3):

$$\begin{aligned} CE_{threshold} = CE_{bn} &= 1, P_{i,j} >= thr \\ &= 0, P_{i,j} < thr \end{aligned} \qquad (3)$$

These are cut values that are applied to the output of the CE model. The CE model is the matrix with assigned CE linkage

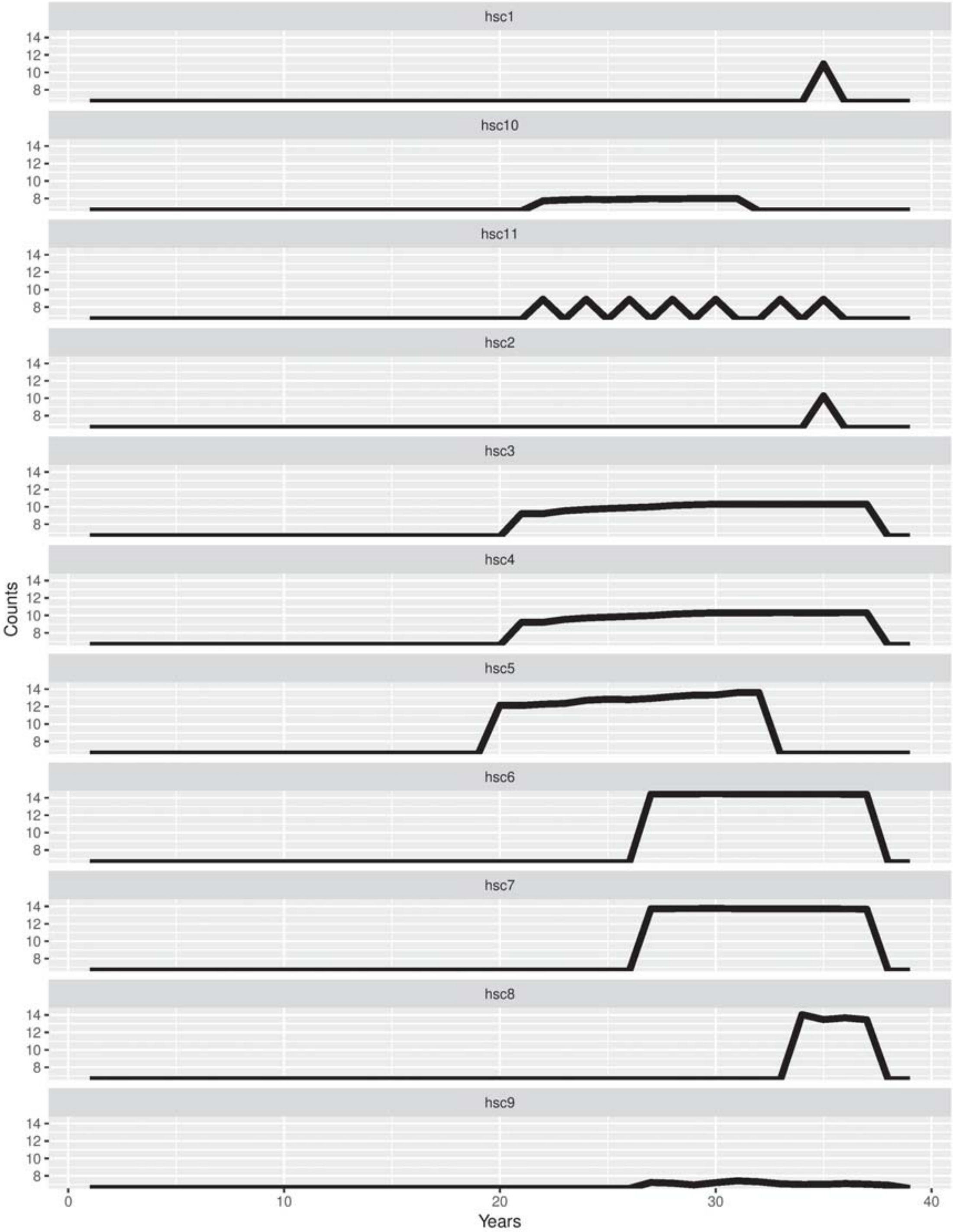


Fig. 1 Logarithms of summed counts (all times series tracked per service) for the 11 packs of H&Sc services over the span of 39 years

Services names, years, attributes and levels break-down ***servicesnames***

IDs	**(HSC)s full names (packs names)**	**Attributes (levels)**	**Years**
1	Self-assessed young people smoking (S1)	**Age**:A1(13:L1,15:L2,All:L3) **Gender**:A2(M:L1,F:L2) **SIMD quintiles**:A3 (Most deprived:L1 ,...,Least deprived:L5) **Smoking behaviour**:A4 (Non.smoker:L1, Occasional smoker:L2, Regular smoker:L3) **Self assessed general health**:A5 (Bad:L1,Fair:L2, Good:L3, Very bad:L4, Very good:L5) **SIMD quintiles**:A6 (Most deprived:L1 ,..., Least deprived:L5)	98-10
2	Smoking and deprivation (S2)	ß **Age**:A1(13:L1, 15:L2, All:L3) **Gender**:A2 (M:L1, F:L2) **Smoking Behaviour**:A3 (Non.smoker-L1, Occasional smoker-L2, Regular smoker-L3) **Self assessed general health**:A4 (Bad:L1, Fair:L2, Good:L3, Very bad-L4, Very good-L5)	98-10
3	Smoking behavior and self-assessement of health (S3)	**Gender**:A2 (M:L1, F:L2), **Weight Category Epidemiological**:A3 ß(Healthy Weight:L1, Obese:L2, Overweight:L3, Overweight...Obese:L4, Underweight:L5)	08-19
4	Epidemiological BMI in Primary 1 Children (S4)	**Gender**:A1(M:L1,F:2), Weight Category ßEpidemiological:A2 (Healthy Weight:L1, Obese:L2, Overweight:L3, Overweight...Obese:L4, Underweight-L5)	05-09
5	Clinical BMI in Primary 1 Children (S5)	**Gender**:A1(M:L1,F:L2,All:L3)	05-09
6	BMI Distribution in Primary 1 Children - Client Group of in Care Home (S6)	**Adults**:A1 (with learning disabilities:L1, Mental health problems:L2, Physical disabilities:L3, All Adults-L4, Older people aged 65 and older:L5, Other groups:L6)	02-11
7	Home sector (S7)	**All sectors**:A1, **Authority and NHS Sector**:A2, **Private sector**:A3, **Voluntary sector**:A4)	07-17
8	Number of general practices - registered patients (S8)	**Type of tenure**:A1 (All:L1, Owned mortgage loan:L2, Owned Outright:L3, Rented:L4), Household Type:A2 (Adults:L1, All:L2, Pensioners:L3, With Children:L4, Age:A3 (16-34:L1,16-64:L2,35-64:L3, 65 and over:L4, All:L5), **Gender**:A4 (M:L1,F:L2,All:L3)), **Limiting long term physical or mental health condition**:A5 (All:L1, Limiting condition:L2, No limiting condition:L3)	96-18
9	Mental wellbeing SSCQ (^{*1}) (S9)	**Value**:A1	05-09
10	Low Birthweight (S10)	ß**Birthweight**:A1 (Live singleton births-L1, Low weight births-L2, Value(all):L3	00-19

Fig. 2 Service packs full names, with attributes and levels breakdowns

probabilities. These are variably cut using P_1 or P_2. Then, the emerging links that appear have one "target" (the effect service on the left) and one or more cause services (on the right). As shown in Fig. 3 the service with (ID=40), as an effect-service, links to five cause-services with confidence above (0.1), then in Fig. 3 (b) the service (ID=8) with threshold (0.1) links to three, in Fig. 3 (c) the service (ID=19) with threshold (0.1) links to five, in Fig. 3 (d), the service (ID=34) with threshold (0.1) links to two others, in Fig. 3 (e) the service (ID=10) with threshold (0.5) links to one, and in Fig. 3 (f) the service (ID=21) with threshold (0.5) links to two. The names of the services are shown on the figures.

B. CE and Cross Correlation

The CE is often confused with Cross Correlation in that well correlated services are taken as CE pairs. As with Cross Correlation it needs to be noted that there is no proved way to cut off (apply threshold) specific CE or Cross Correlation (CC) associations. This is so unless there is a specific prior knowledge (preference) about the solution we are looking for. This is a challenging research question also tackled in other areas as in [39]. The last work states that to best link two data sets we need to check the statistical significance of their connection. That means a high CC still bears a level of significance. Another way is to check if alternative methods (such as the χ^2 test or other tests) suggest the connection or not of two data streams. In this work we had heavy zero padding for time-alignment This raised the levels of CC. Using only CCwould for indicate false CE pairings. Finally, one could improve (refine) the cut-values of the CE connections after tracking (updating), in the long run, the found probabilities of the CE links. The rationale is to check, in the longer term, the validity of such a connection that can stand over time and also face the problem of spurious connections. This is in the sense that spurious connections cannot hold for long.

C. The Hidden or Latent Variables

As it is explained in the equations introduced in [17] the CE linkage between two variables (X,Z) can be explained by other 3^d factors (Y)s (one or more). This link limits the number of spurious connections that cannot satisfy these equations. Co-variant services that link CE pairs are often unobserved or hidden services that may come into play in the background. These (Y)s can link the cause service X with one of the observed outcomes Z. The connections are accompanied by a certain amount of vagueness as to what likely causes what else. Often, this may not be directly interpretable or it can be variably understood. As a result of this one needs to develop hypotheses to approach their validity.

D. CE Modeling

CE, and LR are different areas of ML but can also be interlinked. The CE depends on what is the probability of having the cause. The main effect studied, $\hat{y}_i$ can depend on other side effects, $\hat{x}_i$ that are taking place as co-occurrences with the studied one. The usual easy to model this dependence is by using the Gaussian multivariate model in (5):

$$p(\vec{y_i}/\vec{x_i},\beta,\sigma^2) \approx \frac{1}{\sqrt{\pi}} \times \sigma^{-n/2} \times exp^{\frac{(\vec{x}-\hat{\vec{x}})*(\vec{x}-\hat{\vec{x}})^T}{2}} \tag{4}$$

that for self-prediction becomes $\vec{x_i} = \vec{y_i}$ and $n = 39$. We have $\vec{x_i} \neq \vec{y_i}$ for prediction from other data. That means that to

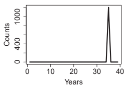
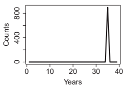
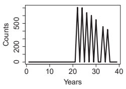
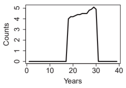
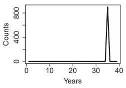
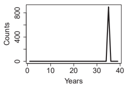
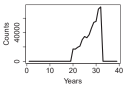
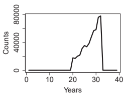
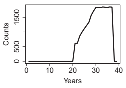
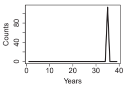
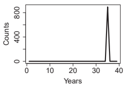
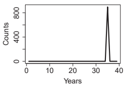
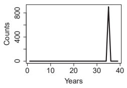
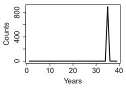
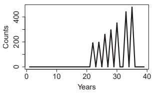
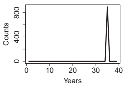
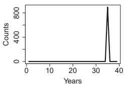
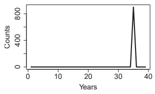
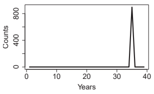
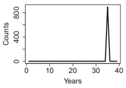
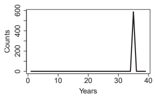
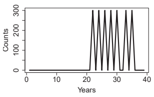
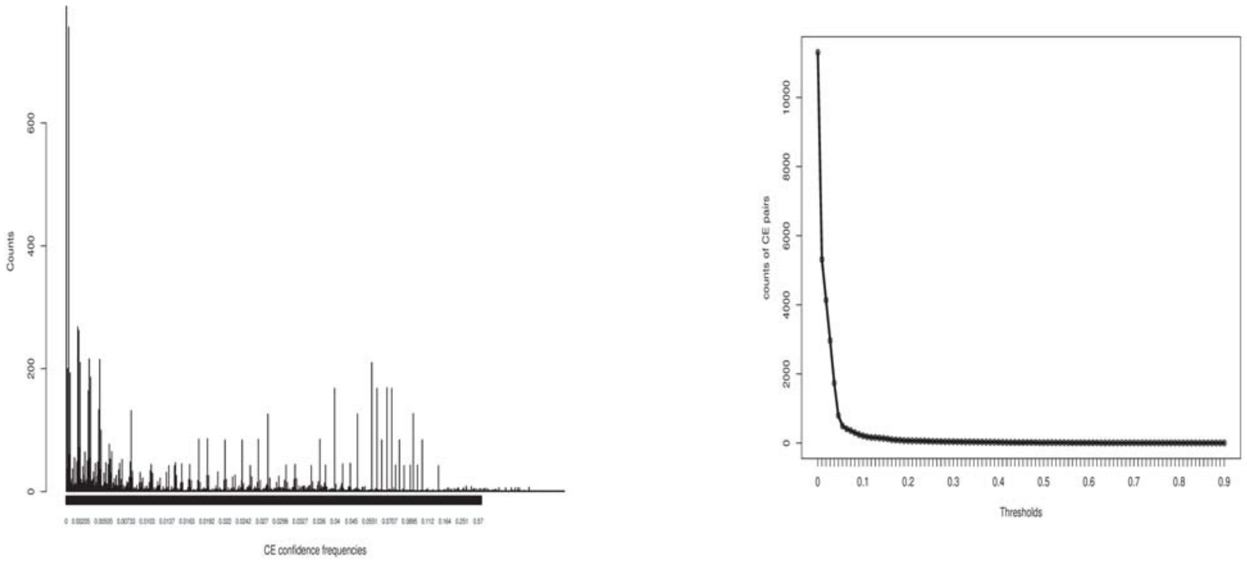
Effect service	Cause services	
 **Effect service (ID=40)** Intensive Home Care . Gender . Male, threshold $>= 0.1$	 Self-assessed young people . Age . 13	 Self-assessed young people . Smoking behaviour . occasional smoker
	 Self-assessed young people. Age . 13-15	 Self-assessed young people . Smoking behaviour . occasional smoker . Age . 13
 **Effect service (ID=8)** Clinical BMI in Primary 1 Children, threshold $= 0.1$	 Self-assessed young people smoking . Gender . All	 Self-assessed young people smoking . Self assessed general.health . Bad
	 Self-assessed young people smoking. Smoking behaviour . Non.smoker	 Self-assessed young people smoking. Age .13
	 Self-assessed young people smoking. Gender . Male	 Smoking prevalence and deprivation SALSUS . Age . 13
 **Effect service (ID=19)** BMI Distribution in Primary1 Children - Client Group of in Care Home, threshold $>= 0.1$	 Self-assessed young people . Age . 13	 Self-assessed young people. Age . All
	 Self-assessed young people. Gender . Female	 Self-assessed young people . Smoking behaviour . Non smoker
 **Effect service (ID=34)** Mental.Wellbeing.SSCQ, threshold=0.5	 Self-assessed young people . Age . 13	 Self-assessed young people. Age . All
 **Effect service (ID=10)** Epidemiological BMI in Primary 1 Children threshold=0.5	 Intensive home care . Age . 15	

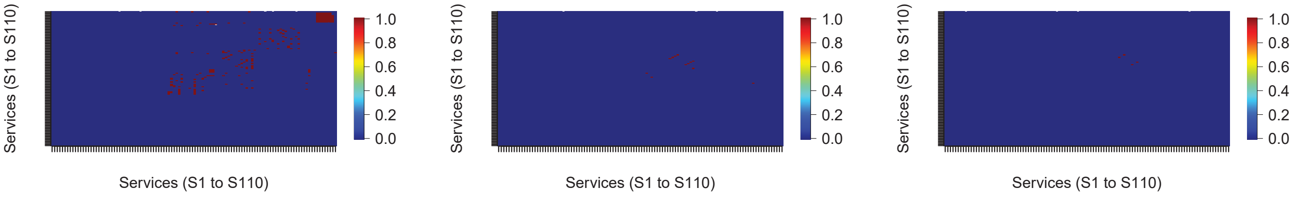
Fig. 3 Indicative CE relationships plots: The left column are the effect-services and the right column one or more cause-services of them



(a)

(b)

Fig. 4 The CE confidence (X-axis) frequencies (Y-axis). There are (1046) different confidences observed. the dropping number of CE paired connections as the confidence threshold increases. The remaining CE links are defined from (3)



(a)

(b)

(c)

Fig. 5 As we change the thresholds for the accepted confidence in CE relationships the number of valid CE pairs falls exponentially (the points kept are the red-colored ones); The axes show the (ID)s of the services that are connected (5a) threshold=0.1, (5b) threshold=0.5, (5c) threshold $(= 0.9)$

define the probability that the level of the demand for a cause service is $p(y(t)) = a$, we would need to use a CE model of the previous samples of both data streams, $(\vec{y}(n-1), \vec{x}(n-1)), n > 1$. The models, $(p(\vec{y_i}/\vec{x_i}, \beta, \sigma^2))$ are the typical LR prediction where the LR parameter β is the one in (2).

The work in [17] very broadly represents the CE models as observed triplets of services of the form in (5):

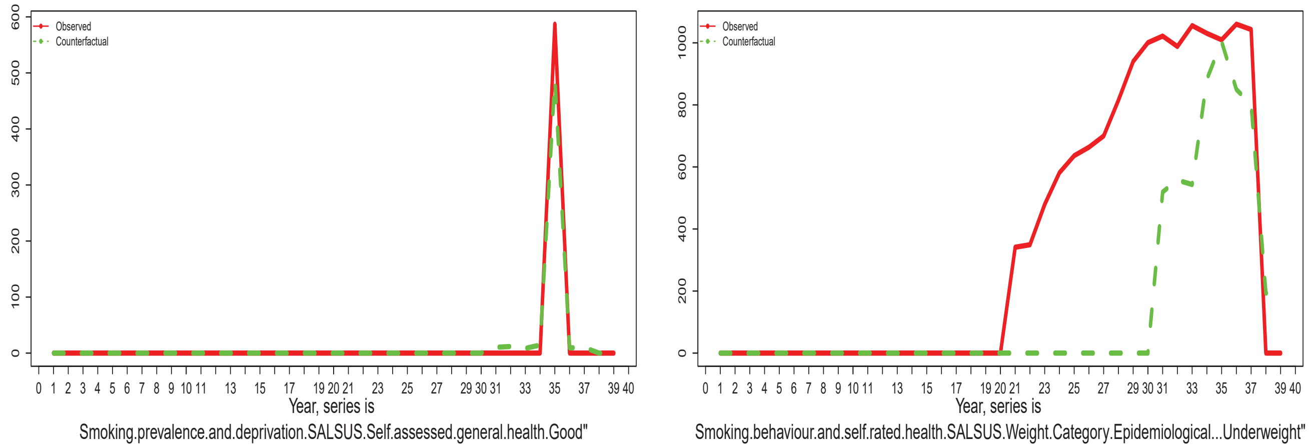
$$\begin{aligned} CE_i = &\{X_i(t), Y_i(t), Z_i(t)\} \\ &i \in [1, 110], t \in [t_1, t_2] \\ &30 <= horizonStart <= t_1 \\ &t_2 <= horizonEnd <= 39 \\ &X, Y, Z \in D \end{aligned} \tag{5}$$

Then, the CE model links the three basic services in the generic relationship: $Z \approx X + Y$. This is a very simplistic version of (2) and implies a cause (X), then an intervention (Y) (that is still another public service) and finally an effect (that is another resulting public service as well, (Z)). The CE model is computed from data from the year (t_1) through the year (t_2) and can link services from the year (t_2) through the end. The duration, d_i is taken here the same for all and is: $d_i = 39 - t_2$.

This also relates to service classification in the sense that a triplet or a quadruplet can form a service cluster in which the target (effect) can be linked to one or more of its co-founders. This can help policymakers to make informed decisions and allows effectiveness in estimating the impacts of applying new measures.

E. Relationship between LR and CE, and Cross Correlation

The relationship between LR and CE is manifold and one case is reported in [40] where the selection of the cause variables can be linked to criteria based on LR. Then, from the CE matrix computed one tried different confidences in the region $[0.1, ..., 0.9]$ to extract the best drivers (causes of the effect service).



(a)

(b)

Fig. 6 Indicative cases of observed and counterfactual data; İn (6a) observed match well counter-facts, in (6b) counter-facts spread more

In the experiments conducted the ARMA model was used as a self-causing method and crashed for large lags. In [41] the ARMA models apply to likely linearly related year series. In [42] a step-wise regression is suggested as a better one over traditional LR or ARMA methods. The step-wise regression is also relevant to a changing CE model discussed here for pairs of services that have their own duration. We may get higher ($CE_{i,j}$)s in (5) when the CE matrix is computed over more years (that is the time $t_2 - t_1$) discussed before. Often a changing model is found empirically using specific ad-hoc ranges for the CE models. An interesting point is made in [43] where it is advocated that CC needs to be coupled with ML to reveal specific relationships across data. A work that can help as well on that matter is in [44], that is, on the use of CC along with Granger causality tests. An interesting point made is that CC is not causality. In the matrix for CE in (5) and in (1) we have taken as granted that the diagonal terms are self-causalities ($X \equiv Z$). These are the most likely CE pairs with ($P_{i,i} = 1$). This means that very high CE can lead to high CC, but not the opposite. That is, a high CC does not necessarily imply a high CE. The distinction of valid CE pairs from spurious CE links can be based on the fact that the CE links are more generic, as an idea, than simple Cross Correlation is. For example, in the seminal work of [45] the CE link is a statistical hypothesis test that is based on more statistics metrics that may include tests such as F-test, χ^2 tests, ANOVA. The causality that can be insured using the ability to predict is then called *'predictive causality* '. The Cross Correlation is not limited to that and it is distinct from the CE links. This is because it is known that when two data streams are well correlated then this does not imply their CE relationship. Indeed, the Granger causality checks how well two data streams (here these are service demand data) x and y can cause (predict) one on the other. To check this one way is to consider to which extend relationships of the type $y_t = \sum_{i=0}^{i=p} a_i * y(t-i) + \sum_{j=0}^{i=q} b_j * x(t-j)$ hold for each of the (x_i), (y_i)s individually. This same principle of confident (a_i)s or (b_i)s is also taken into account in typical LR models. When the t-test and the F-test prove true (that is the hypothesis that x helps to determine y) then these are said to be Granger-test related. Again, the CC is quite broad (not specific) to ensure such a relationship. Moreover, as discussed in [46], the causality relationship between two data streams is contrasted to their spurious correlation. The reasoned CE connection versus the spurious one cannot be determined. A human observer can connect, often arbitrarily, and through a mental process rather than a mathematical process one phenomenon to another phenomenon. This connection very often depends on many circumstances. Many of these circumstances may not be easily recorded as quantitative data.

The implementation of the causality test under the software package (R) is the Wald-type test and the Granger test. The Wald test can check the significance of the contribution of a single cause to an effect one using an estimate of the corresponding error. As the work in [47] discusses there is a time lag (timed information flow) that is part of the Granger test suggesting that a cause needs to precede the effect. The Granger test is well adopted in the econometrics area as in [48] where the expenditures per capita in healthcare is modeled as a function of the GDP. The Granger test, the LR, and the CC enlighten different aspects of the data. These aspects can concern prediction, similarity (CC), CE relationships, and grouping. The LR models can check for likely linear relationships among the data determined by probabilities as in [49]. In [50] LR is used to predict daily patient discharges using 20 patient features and 88 hospital-ward level features and other administrative data. The discharge as an effect can thus be linked to both clinical and other causes.

LR can relate different service attributes to patients' parameters to create cohorts so that similar data can be

Open Science Index, Computer and Systems Engineering Vol:18, No:9, 2024 publications.waset.org/10013811.pdf

mutually predicted. CE links them as CE pairs. LR is not identified by CE (not the same) although both entail combinations of services in different contexts. With LR a service is any data stream $y_i, i \in [1, 39]$ that can be approached by its LR representation across the year span which is called its prediction. CE associates this data stream as the result of past data that in some way caused it to take values. The LR prediction can be part of the CE models.

Typical regression cannot find possible effects on causes since it relates standard predictors (founders) to an impact (effect). It only allows the direct relation without co-founders' mutual-interference. These co-founders can be pre-determined in equations like (2). LR relates by conception one target to more variables while CE can link many to many. CE uses, though, LR to define the best causes in the sense of least squares. The mutual interference of services (how one depends on another one) is accounted for only in linear collinearities which allow for products of such factors (founders or predictors) to be taken into account. In services mutual interference one first considers the impact one onto another (through their product) and then both interact (have an effect on) the outcome. For the simplistic CE model where the cause variable is X, the effect variable is Z, and the co-founding variable is, Y this relationship becomes, $X \Leftarrow Y \leftarrow Z$. This is a relationship that is depicted in [29] where the investigation of new therapies and unobserved treatment effects on volunteers is discussed. A challenge discussed in this work is that the effects of alternative treatments often cannot be determined (systematically observed) due to the heterogeneity of the responses from administering different treatments on the subjects. Even for the same treatment one can observe a large variety of responses (clinical outcomes). Even within the same population onto which a new drug or therapy was administered one may not observe over years the same response. Then, the difference with respect to the response (the outcome or the effect variables) from the control group may change over the years. This method is then called "Difference-in-Differences".

IV. Results and Discussion

Generally speaking, services that relate to mental conditions such as "Mental well being (SSCQ)" proved linearly linked (part of LR relationships as in [43]) to more services such as those related to smoking "Self-assessed young people.Age.13". In the context of CE (confidence around (0.1)) part of these relationships still holds with higher confidence compared to LR (where the confidence found was (0.001), not shown here) or below the confidence (significance level of (0.05)). That means that public service connections using CE models do not exclude their LR connections. The LR is less confident and involves at least 3 other (independent) public services whereas with CE the cause services cannot be more than 4 and usually 2 or 3 with confidence likely around ($P = 0.1$) and less likely around ($P = 0.5$). The relationship of the number of highly confident CE connections to the level of it is shown in Fig. 4b where one can see how the counts of CE pairs per cut value ($\in [0.1, 0.9]$) drops. Another finding from comparing LR with CE (not shown here) is that the effect services in CE models can be rather predictors of other effect services and not of cause services. This may be due to the nature of the data this work examined which used many TS. For example, we had many single data streams (TS of single attribute year series) derived from hospital admission counts per reason of admission. There, the reasons for admission were mainly: (1) alcohol-related data streams (admissions) (2) admission for giving birth (3) smoking-related admission. These cohorts or services were mostly encountered as cause-cohorts or services. The rest of the 110 examined TS were the effects. The effects were more often observed as linked using LR models to other services that were still effect ones but were not as likely observed as cause services. That is, one effect service was found to likely LR predict another effect service with any of the 3 admission reasons as co-predictors.

Also, in the LR models the admission data (TS) were not likely observed as predicted by the non-admission data. In other words, and as expected, the admission reasons were more likely causes of non-admission related data and also could not likely be predicted as they were mainly observed as prediction variables in LR models. This is reasonable since the three reasons for admission (and the related data) are rather mutually independent.

The CE models assigned confidences in the interval $P_{CE} \in [0.001, 0.9939516]$ and with frequencies shown in Fig. 4. There are 1046 different confidences (that is, computed using CE models) and the maximum count is 600 (that is, the most popular probability among the 1046 found). The maximum number of causes per different effect is 109 (=110-1). This is when the effect is not the same as its cause. However, the way that the 109 services can be combined, in theory, to cause the $110-th$ (effect) service is, $\sum_{i=1}^{i=109} \binom{109}{i}$, where, $\binom{109}{i} = \frac{109!}{(109-i)! \times i!}$ since we can have combinations of any number (n) of cause-services, $n \in [1, 109]$ to have an effect one. Hence, 600 seems a reasonable number of paired combinations for CE confidences above (0.1) which is what the maps in Fig. 5a illustrate. The maps in Figs. 5a and 5c relate, using singles paired. That is a single cause on Y-axis to a single effect point on X-axis. There are 110 such effect-service points on the map that result from the still 110 cause services each. The map follows a color mapping of the confidence levels (as per the color scale shown). We can have in theory as many CE maps as the times we can cut-off the confidence level that the CE models yield. That is, we can set a threshold on what to show (above the threshold) where a CE link exists. The three plots show, in Fig. 5a the map produced for a minimum cut-level, ($P > 0.1$). Then, in Fig. 5b the map is for ($P > 0.5$), and finally in Fig. 5c the CE map is for ($P > 0.9$). The exponentially falling curve in Fig. 4b expands on what is shown from Figs. 5a-5c for the case of 100 cut-values, $CutValue \in [0, step = 0.01, 0.9]$. For each of them shown is the number of CE-paired relationships that can be found if one keeps only those that are above the current *cutValue*. The final one (count not shown in Fig. 4b) is 4 (that is four CE pairs only) which is assumed for the (cutValue)s in ($0.8272727(0.827) < cutValue <= 0.9$). That

Open Science Index, Computer and Systems Engineering Vol:18, No:9, 2024 publications.waset.org/10013811.pdf

means that for those values in ($cutValue > 0.5$) one can find only 4 CE pairs of public services that can be linked as CE links. The actual value for > 0.5 is ($cutValue = 0.8272727$). These services, excluding those from one service to itself (that are 3) are: {"BMI Distribution in Primary1 Children - Client Group of in Care Home (older people aged 65 and older)", "BMI Distribution in Primary1 Children - Client Group of in Care Home (All Adults)}, with a probability ($cutValue = 0.8272727 (= 0.827)$).

A. Major Findings

Fig. 6a shows good tracking results in CE pairing and that results (effects) follow the causes. It is also interesting that counter-effects tend to spread more than the observed data. This is mainly observed in Fig. 6b. Figs. 6a and 6b show co-plots of indicative cases of observed and counterfactual pairs (services IDs, #10,#50) that were considered more interesting. Fig. 3 shows the service "Clinical BMI in Primary 1 Children" is linked to five other services like (a.1) "Self-assessed young people smoking . Gender . All" ($P = 0.482$), (a.2) to "Self-assessed young people smoking . Self assessed general.health . Bad" ($P = 0.417$), to (a.3) "Self-assessed young people smoking. Smoking behavior . Non.smoker" ($P = 0.417$), to (a.4) to "Self-assessed young people smoking. Age .13" ($P = 0.405$), (a.5) to "Self-assessed young people smoking. Gender . Male". All these (P)s are above ($P = 0.1$). One can see that services in the H&Sc-pack related to smoking can be taken as a cause of services related to the BMI index in the pack "Primary 1 Children" (children in this age band). Smoking in young people can be indeed related to the newborns indexed by BMI as it is also discussed in [51] in regards to body length (height) as well as in [52] where both factors (that is, early pregnancy and smoking) seem to be related to an increase in the percentage of (LBW)s (Low Birthweight).

Another finding from Fig. 3 is that one cause-service can be variably linked to more effect-services. For example, one can see again in Fig. 3 the dominance of the service (or social cohort) of "Self-assessed young people .Age. 13" that is linked to (1) "Mental.Wellbeing.SSCQ", (2) "Epidemiological BMI in Primary 1 Children", (3) "BMI Distribution in Primary1 Children - Client Group of in Care Home", and so on. This can also imply the existence of co-founders, that is other services that cause these services by affecting other variables that may more directly have an effect on them (for example with higher CE probabilities). The examples in Fig. 3 are quite limited to exploring relationships of type-I ($X \to Y \to Z$) as discussed in the introduction. Another interesting point in Fig. 3 is that most of the services CE relationships link different (H&Sc)s, that is, services that do not belong to the same HSC pack. This helps in planning and in designing policies since one can associate, using CE pairs, different parts of the population (that is, as cohorts) and plan the resources needed.

Fig. 3b shows another row (CE pair) of the CE matrix where "Intensive Home Care.Gender.Male" is linked to (b.1) "Self-assessed young people.Age.13" ($P = 0.168$), to (b.2) "Self-assessed young people Smoking Behavior Occasional smoker" ($P = 0.157$), (b.3) to "Self-assessed young people.Age.13-15" ($P = 0.157$), to "Self-assessed young people.Occasional smoker.1" ($P = 0.157$), to (b.4) "Self-assessed young people.Age.13" ($P = 0.157$). Again, All these (P)s are above ($P = 0.1$). One can see in this CE relationship that smoking habits can also relate to males in need of intensive home care. This cause can also be related to the elderly.

Fig. 3c shows more CE pairs such as "BMI Distribution in Primary1 Children - Client Group of in Care Home" (primary 1 children living in care homes) that as an effect can be linked to (c.1) "Self-assessed young people.Age.13" ($P = 0.541$), to (c.2) "Self-assessed young people.Age.All" ($P = 0.166$), (c.3) to "Self-assessed young people.Gender.Female" ($P = 0.164$), to "Self-assessed young people.Smoking Behaviour.Non smoker" ($P = 0.161$), to (c.4) "Self-assessed young people.Smoking Behaviour.Regular smoker" ($P = 0.102$). As expected, children at the age of 13 or 15 as well as young people of any age or gender who smoke are related, through CE pairs, except that now the difference to Fig. 3c is that we have children raised in care homes by young people and smoking habits are involved with gender as "causes" for them to live in care homes. The age seems more relevant (has higher confidence) to that than smoking habits.

In Fig. 3d one can see CE pairs suggesting that "Number of general practices - registered patients" (patients registered with GPs) as an effect can be linked to (d.1) "Self-assessed young people.Age.13" ($P = 0.108$), to (d.2) "Self-assessed young people.Age.15" ($P = 0.109$). This suggests that young people at the age of 13 or 15 tend to register with GPs with average confidence.

Fig. 3e suggests an interesting connection that people who declared as being in the cohort "Mental Wellbeing . SSCQ" can be the "result" of "Self-assessed young people.Age.13" with high ($P = 0.5$). This suggests that young people aged 13 tend not to have mental problems which is expected as mental problems begin usually after some age.

Another strong CE connection is found in Fig. 3f, where the "Epidemiological BMI in Primary 1 Children" is linked as an effect to: (f.1) "Self-assessed young people.Age.13" (P=0.574) as its cause. This link is stronger and as such it only pairs two services. The justification is roughly the same as the previous ones that link BMI as an index to young people (young people are linked to their BMI index).

Moreover, "BMI Distribution in Primary1 Children - Client Group of in Care Home" is also linked to the same category of services (cohorts) as to "Self-assessed young people.Age.15" ($P = 0.560$) and to "Self-assessed young people.Age.13" ($P = 0.531$) with high chances.

It was also found that with LR connections, on average, a number of factors in the region ([2,6]) were well linearly connected (that is, by using LR models) whereas very few quite confident CE pairs ($P >= 0.5$) were found and with no more than one effect CE-linked to a maximum of two causes. As discussed, the service (cohort) ('HSC#2 . Age . 13') defined in the table in Fig. 2 was found as a cause of at least 3 CE pairs (as per Fig. 3). On top, this service was found as the target of 3-5 independent factors as the dependent service. Similar sizes in LR models and in the region ([2,5]) were also discussed in [53].

Although most (H&Sc)s were related using some LR models only a few were related using CE relationships. Also, as it is shown in Fig. 3, in most cases cause services rather preceded their effect counterpart, but again this is not part of CE pairing definition. As it can be seen in the above figure most effect services were single peaks occurring almost right after a burst of cause services were recorded as taken. This is also in line with the findings from the side-LR analysis (that is, a comparison analysis but not the main analysis) where one saw that a single or a few (2) years records services like "Smoking prevalence among 13 and 15-year-olds in Scotland. Health . Fair" (year:2017) were well LR connected to others. Moreover, it was also found that others with no records after 1997 were well connected as well, or those with a single low attendance before 1997. Most (H&Sc)s or factors that did not have records then were mainly met (observed) as effect ones. Also, those H&Sc factors with only very recent records, i.e., after 2017 and not before like 'Delayed discharges: monthly census. other living conditions< all levels >' did not relate well (few cases).

For inclusion in the CE models, no confidence level (P) was adopted. One simply used a cut value to filter the CE matrix in (5). With linear groups, members would be included (considered as well linearly linked to the same target) if their LR coefficients had low probabilities (low p-value for non-dependence) and the accuracy (RMSE) was kept to an acceptable level ($RMSE >= 0.8$). The probability levels depended on the number of independent H&Sc factors used. The average observed was close to 1 percent and above, 0.8. With LR the accepted probabilities belonged to the interval ([0.001,0.05]). Extreme cases below 0.001 or above 0.05 or with $RMSE \in \{0,1\}$ were an over-fit or a non-fit and were ignored. The $RMSE = 1$ was accepted but not with too low or too high probabilities.

The services reported as more likely effects of other services if one considers as likely CE pairings (one effect, more causes) those for ($P > 0.1$) are: (1) "Primary 1 BMI Distribution . Care Home Sector . Private Sector" (11 causes), (2) "Primary 1 Children.BMI Epidemiological.Weight Category . Clinical...Obese...Severely.Obese" (10 causes), (3) "Intensive Home Care.Age.15" (8 causes), (4) "Intensive home care . Age.All" (8 causes), (5) "Intensive home care . Gender.All" (8 causes). One can see that services related to young people (children) directly (newborns weight), or, indirectly, such as those who live in care homes where there are children and they are parents of them have services addressed to them as causes. The criterion was not the level of the CE confidence attached to them (still $P > 0.1$) as CE pairs but the popularity (count) of the service being a cause to them in the data. As discussed, LR can complement service delivery knowledge gained from CE. For example, for H&Sc pack ("1"), as above, that is an effect of several services one can find most of the services packed as potentially predictable by other services as LR predictors. An example is "Smoking behaviour and self-rated health(SALSUS). Gender. Female", or other-smoking related services. As discussed, the CE and LR models are not the same but use the concept of linearity in a different context. The exact enumeration of how common LR links are with CE links and how they might overlap either as targets or as effects is a level of detail that is not necessary in the context of this work.

In [54] and [55] the number of the factors (predictors) is actually a parameter to adjust which is in our case fixed, i.e., maximum (110). It was found that very good independent factors with the strongest coefficient belonged to the packs: "S20" ("Alcohol-related admissions (stays) or discharges") and "S3" ("Smoking behaviour and self rated health.(SALSUS)") with many likely linear dependencies (that is, below the p.value of (0.05)). Indeed, this can be expected as such causes are dominant for hospital admissions and are at the root of social problems.

Zero padding revealed more relationships and did not limit the results only to common years. For example, the packs: 'S1' (1998-2010) and 'S3' (2008-2019) had very low overlap, and although brought into the same span after zero-padding they were not found well linearly correlated as a pair but they were with other services. An example is 'Headcount of general practice workforce' (S15) with 'Living arrangements for home care clients' (2007-2017) (S14) while 'Alcohol use among young people' is well connected with many but not with 'S3'. The service-ID 'HSC#3' and especially its factor '.type of tenure . owned loan' is a well-modeled (predicted) factor and creates (where it is common) patients categories as can be seen in the same table with probabilities : ({1e-23,0.999,0.968}. Some services connected with a p-value below (0.001) are likely an over-fit.

V. Conclusion and Contributions

The paper discussed how we can relate H&Sc service demands using CE models that can link services as exposure and outcome variables. One contribution of the paper is that this research was conducted on open data published by NHSS that was not attempted before. Another contribution is to the generic discussion on how ML (statistical inference) can be used in policymaking. The linkage of public services, as effects, to other services as their causes are not dealt with in this context. This is also directly related to the financial cost of public services and to cost reduction. This association is also rare to find in this context. Also, the paper presented and discussed a rationale for how the financial risk of new policies can be linked to cost reduction using the concept of spurious connections. The spurious connections entail a risk. This risk is the vagueness of what can be safely linked to what else so that one can, safely or not, spend public money on new policies. When this is not always well known (spur or unknown) or wrong then the decision may have a financial risk. The paper highlighted and discussed this risk using analogous paradigms from other disciplines like social policies, clinical research, and engineering. Another contribution of the paper is that it contrasted concepts such as Cross Correlation, Linear Regression, CE and ARMA prediction in decision-making. The paper stressed that all of these metrics can help decision-making using significance tests. These tests can advise to what extent these metrics can be confident and not confusing. To the best knowledge of the author, this is new in the area and in this very setting of open public services modeling. Furthermore, the paper also linked prediction to inference when it comes to predicting services from other ones to create CE cohorts. As per formulations provided there is a direct link between regression, CE models and also Cross Correlation. Hence, LR can be part of CE equations and also CC is the simplest form of comparing any pair of data. LR is more restrictive than CE is and asks for real overlap of data over time (that is, true co-occurrence). As such it can refine the results of CE and limit the number of data that can be examined as cause-effect pairs. This is also advocated in the mentioned papers. LR can set the final data (as candidate causes for an effect). LR can organize services in linear groups. Then, these groups can be the new refined data space to choose CE pairs from. The paper also gave indicative results to compare LR with CE. The paper also evaluated LR as a separate method for the analysis of data that was distinct from the CE analysis. Thus, LR can have a dual role. It can be used to refine and select candidate causes and effects as linear associates. It was also found that LR links more services than CE does but with less confidence. The dependence of these LR relationships on classification as well as on service settings (attributes) was studied as well. The paper discussed that LR and CE are not the same and differences between them were observed and thoroughly commented. LR links a single target as a predicted demand, to more predictors whereas, CE can link, loosely, more effects to causes. Also CE does not have time limits (co-occurrences) for binding causes to effects and data do not need to overlap in time. The causes can have time-lagged effects or can be linked to latent unobservable data that may link to both the causes and their effects. The looser connection overcomes the restriction to overlap. This was not the case with the explicit overlap of a linear target to its predictors so that the LR coefficients are non-zero values. The CE model was trained using a 6 to 10 year horizon (from roughly 1981 to 2010) that were common to most services from 2011 to 2015. With LR, ARMA the analysis was more brief, since more literature can be found for them compared to CE models. Also there were space limitations to consider. The reason was the need to focus on the deeper presentation of the CE models in public health. CE is not widely used in public services planning and is an under-represented area of research. The paper attempted to show the wide variety of CE models. The LR tests conducted revealed that one can link services using LR within a span of 39 years. Effectively (practically) one can approach (predict) a target from around 2 to 4 other services with a confidence in the region above (0.05) and mostly in the interval $[0.1, 0.4]$.

Another innovation introduced is that public services were analyzed as medical intervention data (causes, interventions, effects). Also the time lag was long (years).

Among the challenges faced were the sparsity of the data over the years, and the scarcity of the data before 2010. This allowed a confined analysis of a span of 6 to 10 years to be carried out. Another benefit of the analysis was to find that services (such as admission to a hospital) that are related to alcohol were a common cause. Also, the results complemented (if not deviated) from the analysis conducted using LR models. It should not be misinterpreted that LR can replace CE in spite of the theoretical relationship found (that is, they entail both somehow dot products of data streams). The LR was only compared to CE in terms of what service(s) can be linked to what other service(s). Indeed, both LR and CE connect services either as predicted/predictors groups (as in LR) or as cause-effect groups (as in CE). Indeed, with LR prediction the involved services came more often from the same H&Sc pack as the predicted one. In the examined CE models the effect services were caused more often by services outside of the H&Sc pack in which the effect belonged. Also, the cause and effect services were more distant in years. The CE models proved that linking services is uncertain and may depend on factors such as the year the data were recorded in. Some H&Sc factors related to self-assessed health status were found widely linked as causes to services related to intensive care services or GPs and to services intended for people with mental health problems. Patient cohorts related to mental well-being and epidemiological birthweight seemed better defined as effects (or well related to) cohorts of self-assessed patients. Such relationships were not found with LR models, that is, well linked services or cohorts from different H&Sc packs. The work also revealed that services that are more common as predictors for other services were related to ‘Alcohol Admissions‘ for example (S20). Also, it was found that home-based services (various services: ‘S11‘, ‘S12‘, ’S14’, etc.) are common reasons for getting admitted to a hospital. Also, it was found that services may expand and differentiate once a patient is originally admitted for one of these reasons. Moreover, a finding was that the HC system has grown around services offered to the elderly or to home-based users as it was seen by the plethora of services offered from a distance and their participation to services groupings. The BMI is an epidemiology index and it was found loosely but widely connected. That is, it was connected to many other serviceswith a low confidence $<= 0.5$) and to many cohorts of young people (ages of 13 and 15) declaring as smokers. Maybe this shows an expansion (breakdown) of the issues relating to low-birthweight, smoking, and to ages that give birth at young ages. Some CE connections are not as obvious are others are. For example, the CE association of young people (aged 13) with those suffering from mental health problems is not obvious and it might imply the existence of lateral dependencies that are not captured. It was also found that GPs workforce could be related to patients self-assessed as being well (SALSUS). Among other findings, low birthweights are related to the people who are offered housing on a voluntary basis and in care homes, and both are linearly related to the patients that are registered with GPs and live in adult-type care homes. These may offer links across the data that may not be expected or even justified. These can be spurious connections but they can also be very useful for policymakers. The merits of using ML is that it can offer out-off-the-box solutions that may offer insights as for hidden data relationships.

Acknowledgment

The author is grateful to PHS and to NHSS for indicating these data to work on. My funding during part of this work came from Abertay's "O924: Integrated Modelling of Health and Social Care Pathways" fund and from "R861: Systems Approaches to Cancer (2016-19)".

References

[1] Simon Bottery et al., A fork in the road: next steps for social care funding reform, The King's Fund. Available at: https : //www.kingsfund.org.uk/publications/fork − road − social − care − funding − reform

[2] Shpitser I., Identification in Causal Models With Hidden Variables, J Soc Fr Statistique (2009). 2020 Jul;161(1):91-119. Epub 2020 Jun 30. PMID: 33240555; PMCID: PMC7685307. Available at: https : //www.ncbi.nlm.nih.gov/pmc/articles/PMC7685307/

[3] Olier, I., Zhan, Y., Liang, X. et al., Causal inference and observational data, BMC Med Res Methodol 23, 227 (2023). DOI: https://doi.org/10.1186/s12874-023-02058-5.Available at: https : //bmcmedresmethodol.biomedcentral.com/ articles/10.1186/s12874 − 023 − 02058 − 5

[4] Maria Ganopoulou, Dimitrios Koparanis, et al., Causal Structure assessment in Health-Related Quality of Life questionnaires, Conference: 35th Panhellenic and 1st International Statistics Conference, Athens, Greece, DOI:10.13140/RG.2.2.32272.58881. Available at: https : //www.researchgate.net/publication/371169040

[5] Salman Afsar, Laura B. McAvoy, Hervé Le Louet, Sec. Advanced Methods in Pharmacovigilance and Pharmacoepidemiology, Front. Drug Saf. Regul., 25 May 2023, Volume 3 – 2023, DOI:https://doi.org/10.3389/fdsfr.2023.1193413. Available at: https : //www.frontiersin.org/articles /10.3389/fdsfr.2023.1193413/full

[6] Joachim P. Sturmberg, James A. Marcum, From cause and effect to causes and effects, 3 March 2017. DOI: https://doi.org/10.1139/er-2016-0109. Available at: https : //onlinelibrary.wiley.com/doi/full/10.1111/jep.13814

[7] Stefano Cucurachi and Sangwon Suh, Cause-effect analysis for sustainable development policy, Environmental Reviews. Available at: https : //cdnsciencepub.com/doi/abs/10.1139/er − 2016 − 0109

[8] European Labour Authority, Measuring the effectiveness of policy approaches and performance of enforcement authorities, November 2022, Available at: https : //www.ela.europa.eu/sites/default/files/ 2023−02/Output− paper − from − plenary − thematic − discussion −measuring − the − effectiveness − of − policy − approaches −and − performance − of − enforcement − authorities − %282022%29.pdf

[9] British Property Federation, The Impact of Rent Control on the Private Rented Sector, May 2023. Available at: https : //bpf.org.uk/media/6296/2023−03−the−impact−of−rent− control−on−the−private−rented−sector−bpf−final.pdf4

[10] David Docquier, Giorgia Di Capua. Reik V. Donner et al. A comparison of two causal methods in the context of climate analyses, EGUsphere, Available at: https : //egusphere.copernicus.org/preprints/2023/egusphere − 2023 − 2212/egusphere − 2023 − 2212.pdf

[11] Wenkai Hu, Jiandong Wang, Fan Yang, Banglei Han, Zhen Wang, Analysis of time-varying cause-effect relations based on qualitative trends and change amplitudes, Computers & Chemical Engineering, Volume 162, 2022, 107813,ISSN 0098-1354, DOI: https://doi.org/10.1016/j.compchemeng.2022.107813 Available at: https : //www.sciencedirect.com/science/ article/abs/pii/S009813542200151X

[12] Kay H. Brodersen, Fabian Galluser, Jim Koehler, Nicolas Remy and Steven L. Scott, Inferring Causal Impact Using Bayesian Structural Time-Series Models, The Annals of Applied Statistics 2015, Vol. 9, No. 1, 247–274. DOI: 10.1214/14-AOAS788.2015. Available at: https : //static.googleusercontent.com/ media/research.google.com/en//pubs/archive/41854.pdf

[13] Oki Oktaviani, Budi Susetyo, Bambang Purwoko Kusumo Bintoro, Risk Management Model using Cause and Effect Analysis in Industrial Building Project, International Journal of Research and Review, Vol.8; Issue: 8; August 2021. DOI: https://doi.org/10.52403/ijrr.20210832 www.ijrrjournal.com ISSN: 2349-9788; P-ISSN: 2454-2237. Available at: https : //www.ijrrjournal.com/ IJRR$_V$ol.8$_I$ssue.8$_A$ug2021/IJRR032.pdf

[14] Arman Oganisian and Jason A. Roy, A Practical Introduction to Bayesian Estimation of Causal Effects: Parametric and Nonparametric Approaches, Stat Med. 2021 Jan 30; 40(2): 518–551. DOI: 10.1002/sim.8761 PMCID: PMC8640942 NIHMSID: NIHMS1755614 PMID: 33015870 Available at: https : //www.ncbi.nlm.nih.gov/pmc/articles/PMC8640942/

[15] Koopmans E, Schiller DC. Understanding Causation in Healthcare: An Introduction to Critical Realism. Qual Health Res. 2022 Jul;32(8-9):1207-1214. DOI: 10.1177/10497323221105737. Epub 2022 Jun 1. PMID: 35649292; PMCID: PMC9350449. Available at: Available at: https : //www.ncbi.nlm.nih.gov/pmc/articles/PMC9350449/

[16] Erik Igelström, Peter Craig, Jim Lewsey, John Lynch, Anna Pearce, Srinivasa Vittal Katikireddi, Causal inference and effect estimation using observational data, BMJ. J Epidemiol Community Health 2022;76:960–966. doi:10.1136/jech-2022-219267. Available at: https : //jech.bmj.com/content/jech/76/11/960.full.pdf

[17] Vidushi Adlakha and Eric Kuo Statistical causal inference methods for observational research in PER: a primer. Available at: https : //arxiv.org/pdf/2305.14558.pdf

[18] Zhouxuan, Kai Zhang, Yashar Talebi et al. Causal Inference for Estimation of Vaccine Effects from Time-to-Event Data. DOI: https://doi.org/10.1101/2023.09.24.23296040 Available at: https : //www.medrxiv.org/content/ 10.1101/2023.09.24.23296040v1.full.pdf

[19] Blöbaum P, Janzing D, Washio T, Shimizu S, Schölkopf B. Analysis of cause-effect inference by comparing regression errors. PeerJ Comput Sci. 2019 Jan 21;5:e169. Doi: 10.7717/peerj-cs.169. PMID: 33816822; PMCID: PMC7924496. Available at: https : //www.ncbi.nlm.nih.gov/pmc/articles/PMC7924496/

[20] Alberto Maydeu-Olivares, Dexin Shi, Amanda Fairchild, Estimating Causal Effects in Linear Regression Models With Observational Data: The Instrumental Variables Regression Model, July 2019, Psychological Methods, 25(2). DOI:10.1037/met0000226. Available at: https : //www.researchgate.net/publication/334402799 _Estimating_causal_effects_in_linear_regression _models_with_observational_data _The_instrumental_variables _regression_model

[21] Gredell (2019), Comparison of Machine Learning Algorithms for Predictive Modelling of Beef Attributes Using Rapid Evaporative Ionization Mass Spectrometry (REIMS) Data, Sci Rep, 9, 5721. Available at: https : //doi.org/10.1038/s41598 − 019 − 40927 − 6

[22] Venkatasubramaniam, A., Mateen, B.A., Shields, B.M. et al., Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity: an application for type 2 diabetes precision medicine. BMC Med Inform Decis Mak 23, 110 (2023). DOI:https : //doi.org/10.1186/s12911 − 023 − 02207 − 2

[23] Bebbington E., Linear regression analysis of Hospital Episode Statistics predicts a large increase in demand for elective hand surgery in England. DOI:01081557=document; 01:09:2016. Available at:https : //www : ncbi : nlm : nih : gov = pmc = articles = PMC4315884 =

[24] Uematsu H, Yamashita K, Kunisawa S, Otsubo T, Imanaka Y., Prediction of pneumonia hospitalization in adults using health checkup data. PLOS ONE. 2017;12(6) : e0180159.

[25] Juang WC, Huang SJ, Huang FD, Cheng PW, Wann SR., Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. BMJ Open. 2017;7(11):e018628. DOI : 10.1136/bmjopen − 2017 − 018628, PMID29196487.

[26] Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A., Multitask learning and benchmarking with clinical time series data. Sci Data. 2019;6(1):96. DOI : 10.1038/s41597 − 019 − 0103 − 9

[27] Muge Capan, Stephen Hoover et al. (2019), Time Series Analysis for Forecasting Hospital Census: Application to the Neonatal Intensive Care Unit. Multitask learning and benchmarking with clinical time series data, Appl Clin Inform.2019, 7(2): pp. 275–289. Available at: https : //dx.doi.org/10.4338%2FACI − 2015 − 09 − RA − 0127

[28] Ambarish Chattopadhyay, José R Zubizarreta, On the implied weights of linear regression for causal inference, Biometrika, Volume 110,

Issue 3, September 2023, Pages 615–629, Available $DOI : https : //doi.org/10.1093/biomet/asac058$

[29] Atul Gupta, Joseph R. Martinez, Amol S. Navathe, Selection and Causal Effects in Voluntary Programs: Bundled Payments in Medicare, NBER Working Paper Series. Available at: $nber.org/system/files/working_papers/w31256/w31256.pdf$

[30] Langton JM2018. Wong ST, Burge F et al. (2015). Population segments as a tool for health care performance reporting: an exploratory study in the Canadian province of British Columbia, BMC Fam Pract. 2020: pp. 21-98. Available at: $DOI : 10.1186/s12875-020-01141-w$

[31] Md Saiful Islam, Md Mahmudul Hasan (2018) A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining, Healthcare, 2018, pp. 6-54. Available at: $https : //www.ncbi.nlm.nih.gov/pmc/articles/PMC6023432/$

[32] Chen Chen, Bin Huang, Michal Kouril, et al., An application programming interface implementing Bayesian approaches for evaluating effect of time-varying treatment with R and Python, Front. Comput. Sci., 16 August 2023, Volume 5 - 2023. Sec. Software. DOI: https://doi.org/10.3389/fcomp.2023.1183380 Available at: $https : //www.frontiersin.org/articles/10.3389/fcomp.2023.1183380/full$

[33] Dimitris Bertsimas, Colin Pawlowski, Ying Daisy Zhuo (2018), From Predictive Methods to Missing Data Imputation: An Optimization Approach, Journal of Machine Learning Research, 18 (2018) 1-39. Available at: $http : //jmlr.org/papers/volume18/17-073/17-073.pdf$

[34] E.M. Mirkes, T.J. Coats, J. Levesley, A.N. Gorban (2018), From Predictive Methods to Missing Data Imputation: An Optimisation Approach, Journal of Machine Learning Research, 18 (2018), pp. 1-39. $DOI : http : //dx.doi.org/10.1016/j.compbiomed.2016.06.004$

[35] deRooij M. (2018), Transitional modelling of experimental longitudinal data with missing values, Adv Data AnalClassif, 12, pp. 107–130. Available at: $https : //link.springer.com/article/10.1007/s11634-015-0226-6$

[36] Scottish Government, statistics.gov.scot Available at: $https : //statistics.gov.scot/data_home$

[37] Public health Scotland. Data and intelligence (2020). A – Z Subject Index, 2020. Available at: $https : //www.isdscotland.org/A-to-Z-index/index.asp.$

[38] Scottish Government(2019). Statistics Service Health and Social Care Data. Available at: $https : //statistics.gov.scot/datahome$

[39] Aristotelis Koskinas, Eleni Zaharopoulou, George Pouliasis, Ilias Deligiannis, ’Estimating the Statistical Significance of Cross–Correlations between Hydroclimatic Processes in the Presence of Long–Range Dependence’. Available at: $https : //www.itia.ntua.gr/el/getfile/2234/1/documents/earth-03-00059-v3.pdf$

[40] Fan Chao, Guang Yu, Causal inference using regression-based statistical control: Confusion in Econometrics, Journal of Data and Information Science 8(1):21-28 DOI:10.2478/jdis-2023-0006. Available: $https : //www.researchgate.net/publication/368691749_Causal\_inference\_using\_regression-based\_statistical\_control\_Confusion\_in\_Econometrics$ 23 May 2023. arXiv:2305.14558v1 [stat.ME]

[41] Bui C, Pham N, Vo A, Tran A, Nguyen A, Le T., Time series forecasting for healthcare diagnosis and prognostics with the focus on cardiovascular diseases. IFMBE Proc, 6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6); BME. 2017;63:138.

[42] Liew BXW, Peolsson A, Rugamer D, Wibault J, Löfgren H, Dedering A et al. (2020), Clinical predictive modelling of post-surgical recovery in individuals with cervical radiculopathy: a machine learning approach. Sci Rep. 2020;10(1):16782. $DOI : 10.1038/s41598-020-73740-7$

[43] Dunsmuir WT., Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: the importance of constructing transfer function autoregressive models. Behav Res Methods. 2019;48(2),2016,783-802. $DOI : 10.3758/s13428-015-0611-2$

[44] Damos, P., Using multivariate cross correlations, ”Granger causality and graphical models to quantify spatiotemporal synchronization and causality between pest populations”. BMC Ecol 16, 33 (2016). $DOI : https : //doi.org/10.1186/s12898-016-0087-7$

[45] Granger, C. W. J. (1969)., Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica. 37 (3): 424–438. doi:10.2307/1912791. JSTOR 1912791.

[46] J. Pearl and D. Mackenzie, The book of why: the new science of cause and effect (Basic books, 2018).

[47] Patrick A. Stokes, Patrick L. Purdon, A study of problems encountered in Granger causality analysis from a neuroscience perspective, PNAS, August 4, 2017, 114 (34) E7063-E7072, DOI: $https : //doi.org/10.1073/pnas.1704663114$

[48] Gülnur İlgün, Murat Konca, and Seda Sönmez, The Granger Causality Between Health Expenditure and Gross Domestic Product in OECD Countries (June 30, 2022). Journal of Health Management, Volume 24, Issue 3 DOI:$https : //doi.org/10.1177/09720634221109306$

[49] Skiera B, Reiner J., Regression analysis, Homburg, Christian, Klarmann, martin. In: Vomberg A, editor. Handbook of market research. DOI:$https : //doi : org = 10 : 1007 = 978-3-319-05542-8_17-1; 2018.$

[50] Shivapratap Gopakumar, Shivapratap Gopakumar, Truyen Tran, Wei Luo, Dinh Phung, Forecasting Daily Patient Outflow From a Ward Having No Real-Time Clinical Data, July 2016. JMIR Medical Informatics 4(3):e25. DOI:10.2196/medinform. 5650. Available at: $https : //www.researchgate.net/profile/Truyen-Tran-2/publication/305522446\_Forecasting\_Daily\_Patient\_Outflow\_From\_a\_Ward\_Having\_No\_Real-Time\_Clinical\_Data/links/57c2c0f408ae2f5eb33917f3/Forecasting-Daily-Patient-Outflow-From-a-Ward-Having-No-Real-Time-Clinical-Data.pdf$

[51] Chan DL, Sullivan EA., Teenage smoking in pregnancy and birthweight: a population study, 2001-2004. Med J Aust. 2008 Apr 14 7;188(7):392-6.$DOI : 10.5694/j.1326-5377.2008.tb01682.x.$. PMID:18393741

[52] Rumrich I, Vähäkangas K, Viluksela M, et al., Effects of maternal smoking on body size and proportions at birth: a register-based cohort study of 1.4 million births, BMJ Open 2020;10:e033465. $doi : 10.1136/bmjopen-2019-033465$

[53] Yang C, Delcher C, Shenkman E et al., Expenditure variations analysis using residuals for identifying high health care utilizers in a state Medicaid program. BMC Med Inform Decis Mak. 2019. Available at: $https : //bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0870-4$

[54] Boelaert, Julien & Ollion, Etienne. (2018)., The Great Regression. Machine Learning, Econometrics, and the Future of Quantitative Social Sciences. Revue française de sociologie. $DOI : 59.10.3917/rfs.593.0475.$

[55] Marno Verbeek (2017), A Guide to Modern Econometrics, 5th edition, Wiley, New Jersey, 2017. str. 520