Using Linked Data to evaluate the impact of Research and Development in Europe: a Structural Equation Model

Amrapali Zaveri[1], Joao Ricardo Nickenig Vissoci[2,4], Cinzia Daraio[3], and Ricardo Pietrobon[4]

[1] University of Leipzig, Institute of Computer Science, AKSW Group, Augustusplatz 10, D-04009 Leipzig, Germany
zaveri@informatik.uni-leipzig.de, http://aksw.org
[2] Medicine Department, Faculdade Ingá, 6114, PR 317 road, Maringá/PR, Brazil - 87035-510
[3] University of Rome "La Sapienza", Department of Computer, Control and Management Engineering, Via Ariosto, 25 - I - 00185, Roma - Italy
daraio@dis.uniroma1.it
[4] Duke University Medical Center, Durham, NC, USA
{jnv4,rpietro}@duke.edu

Abstract. Europe has a high impact on the global biomedical literature, having contributed with a growing number of research articles and a significant citation impact. However, the impact of research and development generated by European countries on economic, educational and healthcare performance is poorly understood. The recent Linking Open Data (LOD) project has made a lot of data sources publicly available and in human-readable formats. In this paper, we demonstrate the utility of LOD in assessing the impact of Research and Development (R&D) on the economic, education and healthcare performance in Europe. We extract relevant variables from two LOD datasets, namely World Bank and Eurostat. We analyze the data for 20 out of the 27 European countries over a span of 10 years (1999 to 2009). We use a Structural Equation Modeling (SEM) approach to quantify the impact of R&D on the different measures. We perform different exploratory and confirmatory factorial analysis evaluations which gives rise to four latent variables that are included in the model: (i) Research and Development (R&D), (ii) Economic Performance (EcoP), (iii) Educational Performance (EduP), (iv) Healthcare performance (HcareP) of the European countries. Our results indicate the importance of R&D to the overall development of the European educational and healthcare performance (directly) and economic performance (indirectly). The results also shows the practical applicability of LOD to estimate this impact.

1 Introduction

Basic research is a crucial important driver for innovation, economic progress and social welfare [2,15]. Scientific production concerns especially basic research, but the results which are generated are not only long-term ones but produce spillovers that have short

and medium term effects on industrial innovation [24]. Europe, as a whole, has a high impact on the global biomedical literature, having contributed with a growing number of articles ($210,433$ publications in public health research [23]) and a significant citation impact [22]. The impact of Europe on broader healthcare and social welfare issues, however, is poorly understood. Although the credit goes to the university research for economic impact, there is no consensus on how to measure it [4].

There have been previous projects that focus on measuring similar impact [6,9]. However, these methods have not scaled up to the challenges of multidimensional assessments that is required to measure the overall impact of research and development on healthcare quality. Moreover, the datasets lack in openness, dynamicity and coverage. Measuring this impact poses a challenging endeavor which involves the identification, gathering and analyzing of diverse data. The recent Linking Open Data (LOD) project offers the possibility to access a large number of datasets in various domains[5], which it possible to quickly extend the breadth of the traditional methods to measure this impact. Extensions include not only measures of overall impact on healthcare, but also indicators, all mediated through measures of research and development. Thus, the objective of this paper is twofold: (i) show the feasibility and the usefulness of combining different LOD data-sources to assess the impact of research and development (R&D) on the economic, educational and healthcare performance, specifically in European countries and (ii) employ a structural equation modeling approach to assess this impact.

Therefore, we retrieved relevant data from two socio-economic datasets already available as LOD – *World Bank* and *Eurostat*. We applied a Structural Equation Modeling (SEM) approach to combine the variables and quantify the impact based on different measures such as economic, educational and healthcare. Performing different exploratory and confirmatory factor analysis model evaluations gave rise to four latent variables which were included in the model: (i) Research and Development (R&D), (ii) Economic Performance (EcoP), (iii) Educational Performance (EduP), (iv) Healthcare performance (HcareP) of the European countries. The results indicate the importance of R&D to the overall development of the european educational as well as healthcare systems (directly) and economic performance (indirectly) and also shows the practical applicability of LOD in determining this impact.

2 Methodology

In this section, we first describe the extraction process of a series of research and development, economic, education, and healthcare-related variables through the use of Semantic Web technologies (Section 2.1). We then describe initial data analysis performed on the dataset (Section 2.2) followed by the theoretical framework of the model (Section 2.3). Thereafter, the Structural Equation Modeling (SEM) approach comprising of two steps is described in Section 2.4.

2.1 Data sources

As the first step, we identified two LOD datasets – the *World Bank* and *Eurostat* – from which all the variables related to healthcare, economic and educational performance

[5] `http://lod-cloud.net`

were extracted. Next, we excluded those variables which showed low data quality, that is mainly those that contained missing values. Thereafter, during the model building phase, the variables that did not covariate with other variables were omitted resulting in 18 variables that were ultimately included. The respective variables from each of these datasets and the extraction steps are detailed in this section. Even though traditional methods of extracting data directly from the datasets website (i.e. manually) was an option, LOD provided the advantage of identifying relevant datasets as well as specifying and retrieving data easily using SPARQL. That is, with the data available in a single standardized structured format (RDF) and with the well supported query mechanism (SPARQL), retrieving the relevant data, was less time consuming. Additionally, this reduced the effort of issuing a different query for each dataset along with the data comparisons especially vis-a-vis differing units (and strategies to normalize these). The standardized RDF format also saved the effort of manually handling and combining different data sheets. Most importantly, extracting data using SPARQL allowed us to work with a large amount of data. In this paper we only deal with 18 variables, however LOD can help in gathering larger quantities of variables in the future and from a larger number of different datasets.

World Bank. The World Bank[6] is an international financial institution that collects and processes large amount of data on the basis of economic models and makes them openly available[7]. The World Bank data has been converted and published as LOD. In particular, the World Development Indicators, which present the most current and accurate global development data accessible, are available as RDF[8].

Variables. From the entire list of the World Development indicators, we specifically chose the following indicators which helped leverage our model:

- *Adolescent fertility rate* (Hcare4), which reports the number of births per $1,000$ women aged $15-19$.
- *Birth rate* (GH1), which indicates the crude birth rate i.e. the number of live births occurring during the year per $1,000$ population. This indicator is estimated at the middle of the year.
- *Death rate* (Hcare1), which is the number of crude deaths occurring during the year per $1,000$ population also estimated at the middle of the year.
- *Fertility rate* (GH2), which is the total number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with the current age-specific fertility rates.
- *GPD*, (Gross Domestic Product) (EcoP1), which is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the product. Data is represented in U.S. dollars.
- *Health expenditure public (% of total health expenditure)* (RD3) reports the recurrent and capital spending from government (both central as well as local) budgets, external borrowings and grants and social (or compulsory) health insurance funds. The total health expenditure is the sum of public and private health expenditure,

[6] `http://www.worldbank.org/`
[7] `http://data.worldbank.org/`
[8] `http://worldbank.270a.info/classification/indicator.html`

which covers the provision of health services as mentioned in the indicator "Health expenditure per capita".

- *High-technology exports (% of manufactured exports)* (EcoP2) are the products with high R&D intensity, such as in aerospace, computers, pharmaceuticals, scientific instruments, and electrical machinery.
- *Immunization, DPT (% of children ages* 12 - 23 *months)* (Immu1) measures the percentage of children between the ages of 12 - 23 months who have received vaccinations before 12 months or at any time before the survey. After receiving three doses of vaccine, a child is considered adequately immunized against diphtheria, pertussis (or whooping cough), and tetanus (DPT).
- *Immunization, measles (% of children ages* 12 - 23 *months)* (Immu2) measures the percentage of children between the ages of 12 - 23 months who have received vaccinations before 12 months or at any time before the survey. After receiving one dose of vaccine, a child is considered adequately immunized against measles.
- *Incidence of tuberculosis (per* $100,000$ *people)* (Hcare2) is the estimated number of new pulmonary, smear positive and extra-pulmonary tuberculosis cases which also includes patients with HIV.
- *Mortality rate, infant (per* $1,000$ *live births)* (Hcare3) is the number of infants dying before reaching one year of age, per $1,000$ live births in a given year.
- *Public spending on education, total (% of government expenditure)* (EduP3) reports the total public education expenditure (current and capital) expressed as a percentage of total government expenditure for all sectors in a given financial year. This public expenditure on education includes the government spending on educational institutions, both public and private, education administration and subsidies for private entities such as students or households etc.
- *Research and development expenditure (% of GDP)* (RD2) reports the expenditures for research and development which are the current and capital expenditures i.e. both public and private on creative work undertaken systematically to increase knowledge. This work includes knowledge of humanity, culture, society as well as the use of knowledge for new applications. R&D covers basic, applied and experimental research and development.
- *Researchers in R&D (per million)* (RD1) reports the number of professionals engaged in the conception or creation of new knowledge, products, processes, methods, or systems and in the management of the projects concerned.

Data extraction. The World Bank data is converted to RDF and available at the SPARQL endpoint `http://worldbank.270a.info/sparql`. In order to extract the variables mentioned above, we queried the endpoint using SPARQL[9]. An example of a SPARQL query for extraction data of the World Bank indicator "Health expenditure per capita" is shown in Listing 1.1. For each of the indicators, the indicator code[10] (e.g. SH.XPD.PCAP) was replaced accordingly. Also, the query was modified to filter results only pertaining to the years $1999 - 2009$ and for the specific European countries[11].

[9] `http://www.w3.org/TR/rdf-sparql-query/`

[10] Obtained from `http://worldbank.270a.info/classification/indicator.html`

[11] Omitted from the Listing 1.1 due to lack of space.

```
PREFIX g-indicators: <http://worldbank.270a.info/graph/world-development-
    indicators>
PREFIX property: <http://worldbank.270a.info/property/>
PREFIX g-meta: <http://worldbank.270a.info/graph/meta>
SELECT ?label ?obsValue  (?refPeriodURI AS ?year)
WHERE {
   GRAPH g-indicators:{
         ?observationURI property:indicator indicator:SH.XPD.PCAP ;
                         sdmx-dimension:refArea ?refAreaURI ;
                         sdmx-dimension:refPeriod ?refPeriodURI ;
                         sdmx-measure:obsValue ?obsValue . }
   GRAPH g-meta:{
         ?refAreaURI a dbo:Country .
         ?refAreaURI skos:prefLabel ?label . }
      } ORDERBY ?label ?refPeriodURI
```

Listing 1.1: Extraction of the data of the World Bank indicator "Health expenditure per capita" using a SPARQL query against the World Bank SPARQL endpoint. Certain prefixes omitted due to lack of space but can be resolved using http://prefix.cc.

Eurostat. Eurostat[12] is the statistical office of the European Union (EU), which provides statistical information to the institutions of the EU in order to to promote the harmonization of statistical methods across its member states and candidates for accession as well as European Free Trade Association (EFTA) countries. Eurostat provides statistical data about various topics such as economy and finance, science and technology, industry, trade and services, population and social conditions etc.[13]. The Eurostat dataset is also converted as part of LOD and is made available at `http://eurostat.linked-statistics.org/`.

Variables. From the entire list of different statistics that is provided by Eurostat we choose the following relevant ones:

- *Annual expenditure on public and private educational institutions per pupil/student* (EduP1), which measures the amount which the central, regional and local levels of government, private households, religious institutions and firms spend per pupil/student. It includes expenditure for personnel, other current and capital expenditure.
- *Biotechnology patent applications to the EPO by priority year, country and metropolitan regions* (EduP4) reports the number of patent applications specifically pertaining towards biotechnology to the European Patent Office (EPO). The data is organized according to the year, country and metropolitan region.
- *Economic active population by sex, age and NUTS 2 regions* (EcoP3) comprises of employed as well as unemployed persons and this indicator reports the number of such persons according to sex and age. NUTS refers to the Nomenclature of Territorial Units for Statistics, which divides the economic territory of the EU into different regions, the NUTS 3 referring to the region for the application of regional policies.
- *Financial aid to students* (EduP2) reports the amount of financial aid provided to pupils and students reported as a percentage of the total public expenditure on education for all levels of education combined.

[12] `http://ec.europa.eu/eurostat`

[13] `http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database`

Data extraction. The Eurostat data is converted to RDF and available at the SPARQL endpoint `http://eurostat.linked-statistics.org/sparql`. Similar to the extraction of data from World Bank, we used SPARQL to retrieve data for the variables belonging to the Eurostat dataset. An example of a SPARQL query for extraction data of the Eurostat indicator "Biotechnology patent applications to the EPO by priority year at the national level" is shown in Listing 1.2. For each of the indicators, the dataset code (e.g. pat_ep_nbio) is replaced accordingly. However, this SPARQL query gave rise to three values for each country per year for three units:

- `http://eurostat.linked-statistics.org/data/pat_ep_nbio#A,BIO,MIO_ACT,AT,1999`
- `http://eurostat.linked-statistics.org/data/pat_ep_nbio#A,BIO,MIO_HAB,AT,1999`
- `http://eurostat.linked-statistics.org/data/pat_ep_nbio#A,BIO,NB_TOT,AT,1999`

These three values differ in the way in which the value is calculated[14] i.e. the number of inhabitants per million, in this example. The "A,BIO,MIO_ACT,ACT,1999" was chosen and included as `?unit` in SPARQL query. Additionally, the query was modified to filter results only pertaining to the years $1999 - 2009$ and for the specific European countries[15].

```
PREFIX ns: <http://eurostat.linked-statistics.org/property#>
PREFIX dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>
SELECT ?country ?observation ?year
FROM <http://eurostat.org/>
WHERE
 { ?val   qb:dataSet              dataset:pat_ep_nbio ;
          sdmx-measure:obsValue   ?observation ;
          ns:geo                  ?country ;
          ns:unit                 ?unit ;
          dimension:timePeriod    ?year .
   FILTER (regex(?unit, "MIO_ACT"))
 } ORDER BY ?country
```

Listing 1.2: Extraction of the data of the Eurostat indicator "Biotechnology patent applications to the EPO by priority year at the national level" using a SPARQL query against the Eurostat SPARQL endpoint. Certain prefixes omitted due to lack of space but can be resolved using http://prefix.cc.

2.2 Exploratory Data Analysis

Performing exploratory data analysis is essential to analyze the quality of the data to detect problematic variables, missing values, outliers and other descriptive information about the data to be used. All the information from each of the 18 variables about each of the 27 European countries covering a span of 10 years (199 to 2009) was retrieved and analyzed. The exploratory data analysis resulted in exclusion of seven countries due to missing information in the 10 year span. Included countries were: Austria, Belgium, Bulgaria, Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland,

[14] `http://timetric.com/search/?q=guadeloupe&n=30&p=10`

[15] Country codes: `http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:Country_codes`. Omitted from the Listing 1.1 due to lack of space.

Italy, Latvia, Lithuania, Netherlands, Poland, Portugal, Slovenia, Spain, Sweden and the United Kingdom.

In particular, when there was a pattern of missing data (lacking the last three years, for instance) the country was excluded from the sample. Otherwise, if no pattern was found a multiple imputation method was applied to deal with the incompleteness [27]. Normality distribution was assessed through the Anderson-Darling normality test [12] to detect oscillations in the gaussian distribution in order to adjust the analytical methods to the appropriate distribution. We used the Mahalonobis distance [29] to identify univariate and multivariate outliers and the Mardia coefficient and multivariate kurtosis to identify multivariate normality [21]. Either univariate or multivariate analysis of the outliers or normality distribution are determinant to define which underlying methods (for e.g. extraction) will be applied in the factor analytical process.

2.3 Models' theoretical framework

The model was initially conceived in order to assess the predictor role of the latent variable[16] Research and Development on the European countries' Economy, Education and Healthcare. Specifically, we hypothesized that *R&D would have a direct effect over the Economy (GDP), Educational status (public spending on education), General Health indicators (birth rate, death rate), Health Outcomes (death rate) including Immunization efforts.* The relation between these variables has been separately reported in a number of studies [4,9,13,18]. Therefore, we gathered a core set of variables (as described in Section 2.1) related to each of these factors which represented the situation affecting Europe. However, for our initial model we only kept the data displayed at the country level, excluding other (although interesting) information at regional level.

2.4 Structural equation modeling

Structural Equation Modeling (SEM) [17,20] is a method that has been used in health sciences [14] economic research [3] as well as education [8] to model causal relations among latent and observed variables. This method evaluates the relation between latent variables. For example, in this study we argue that the general concept of economic performance is only possible to explain through a latent variable specified by other observed variables such as GPD, average income etc. Hence, we used SEM to test the outlined hypothesis of a conceptual model based on the effect of R&D on the economic, educational and healthcare situation of European countries.

Our SEM was tested by the jigsaw method [5]. This procedure expects the adequacy of the measurement variables (latent variables that will enter the model) into isolated confirmatory factor analysis models. By doing this measurement before the structural equations we are able to define the model's identification with the latent variables before testing. Therefore a two step strategy is defined to design a SEM [20]: (1) Specify the latent variables through a sequence of Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) in a way that an EFA is performed to detect latent factors and CFA to confirm its structure. If the latent structure does not show adequate indicators then re-evaluate the EFA by making a sequence of EFA-CFA-EFA-CFA until

[16] Latent variables are those that cannot be measured directly, but is an underlying concept involving other observed variables (variables measured directly).

a adequate measurement model is obtained; (2) Specify and identify the SEM based on the models developed previously by inserting each variable one at a time. The steps we followed are:

Step 1. The first step in a EFA analysis is to define the extraction[17] and rotation methods[18]. Once the non-normal distribution of the data was detected, EFA was performed with Principal Axis extraction method which fits this data distribution better. A Promax (Oblique) rotation was performed because we believed that the latent variables would be correlated [7]. The obtained factor loading values[19] above 0.30 were considered acceptable. Models developed by EFA were then tested through CFA sequentially until an adequate model was obtained.

CFA procedure evaluated the model adjustment and adequacy through fitness indicators, factor loadings and individual item reliability. Weighted Least Square was the estimation method used, due to non-normal multivariate normality that was obtained. The indicators used to assess the fitness of the model were: (i) chi-square, (ii) Comparative Fit Index (CFI) (values superior to 0.90 are accepted as adequate fit and 0.95 as good fit); (iii) Tucker-Lewis Index (TLI) (acceptable fit with values superior to 0.90); (iv) Root Mean Square Error of Approximation (RMSEA) (values inferior to 0.08 are considered as acceptable fit and 0.05 as a adequate fit); (v) Akaike Information Criteria/Bayesian Information Criteria (AIC/BIC) (lower values indicate better model when compared to others) [28,19].

Step 2. SEM was applied to test the hypothetical model using the same indicators as described in the measurement model evaluation (Step 1) as well as factor loadings and individual item reliability. The path coefficients were interpreted as: small effect for loadings <0.10, medium effect for loading until 0.30 and high effect for loadings >0.50 [20]. Data analysis was performed through R Language Statistical Software 3.0 version[16], with the specific SEM analysis developed with the sem package [10].

3 Results

From the original 27 countries that entered the samples, several were excluded due to data incompleteness (cf. Section 2.2). Thus, the final sample was constituted by 20 countries. A total of 18 variables (Table 1) were included in the analysis. In this section, we first describe the results of constructing the parts of the model, in particular, determining the latent and observed variables (Section 3.1). Then, we describe the process of choosing the best fit for the model to reach the best possible adequacy and theoretical reasoning (Section 3.2).

3.1 Latent and observed variables

The first step in constructing the SEM was to choose the latent variables that are relevant for the hypothesis. All the variables extracted from the two LOD datasets were

[17] Extraction method is the statistical approach applied to extract the amount of variance of the data that is shared by the variables revealing the latent constructs.

[18] Rotation technique is used to clarify which variables load into each latent construct.

[19] Factor loading is a metric that indicates the amount of contribution of that specific factor to explain the variance in the observed variable.

Table 1: Descriptions and abbreviations used for each of the 18 observed variables belonging to each of the latent variables of the SEM.

Latent variables	Observed variables	Abbreviation
Research and Development (RD)	Researchers in R&D (per million)	RD1
	Research and development expenditure (% of GDP)	RD2
	Health expenditure public (% of total health expenditure)	RD3
Economic Performance (EcoP)	GDP	EcoP1
	High-technology exports (% of manufactured exports)	EcoP2
	Economic active population by sex, age and NUTS 2 regions	EcoP3
Educational Performance (EduP)	Annual expenditure on public and private educational institutions per pupil/student	EduP1
	Financial aid to students	EduP2
	Public spending on education, total (% of government expenditure)	EduP3
	Biotechnology patent applications to the EPO by priority year, country and metropolitan regions	EduP4
General Healthcare (GH)	Birth rate	GH1
	Fertility rate	GH2
Health Outcomes (Hcare)	Death rate	Hcare1
	Incidence of Tuberculosis (per $100,000$ people)	Hcare2
	Mortality rate, infant (per $1,000$ live births)	Hcare3
	Adolescent fertility rate	Hcare4
Immunization (Immu)	Immunization, DPT (% of children ages $12-23$ months)	Immu1
	Immunization, measles (% of children ages $12-23$ months)	Immu2

conceptualized theoretically as parts of the constructs of the model's theoretical framework. However, in order to develop a latent variable we must assess how the variances of each variables relates to the existence of an underlying latent factor. Therefore, we applied a set of EFAs and CFAs to reach the best possible factor structure to apply to the model [19].

Eigen values and screeplot analysis pointed out for the possibility of four to seven latent factors. Therefore, different EFAs were applied to test for four, five, six and seven

factors structures. The six factor model solution showed better indicators explaining 82% of a variance of the variables in the dataset. However, commonalities indicated problems with the variables RD3 and EduP3 with values inferior to 0.50, which meant that these variables were not contributing enough to the latent factor structure specification. Nevertheless, we decided to test how the factor structures would fit in the CFA models with and without both variables.

CFA models were developed for all the six latent constructs, however we had problems in finding convergence in two of the them (GH and Immu). Both latent constructs had only two observed variables specifying them, which might explain the lack of convergence since the recommendation is to have at least three [19]. Also, variance explanation for both models was lower than 10%. Therefore, we decided to exclude four observed variables (i) birth rate (GH1) and (ii) fertility rate (GH2) which constituted the latent variable General Health (GH), and (iii) immunization DPT (Immu1) and (iv) immunization measles (Immu2), which formed the latent variable Immunization (Immu). Furthermore, the variables health expenditure public (RD3) and public spending on education (EduP3) were influencing the adequacy of their respective CFA models, therefore we excluded them from the analysis. A second EFA was performed for the whole dataset, now without the four variables mentioned before. This time, eigen values and screeplot suggested to a three or four factors possibility as shown in Figure 1.

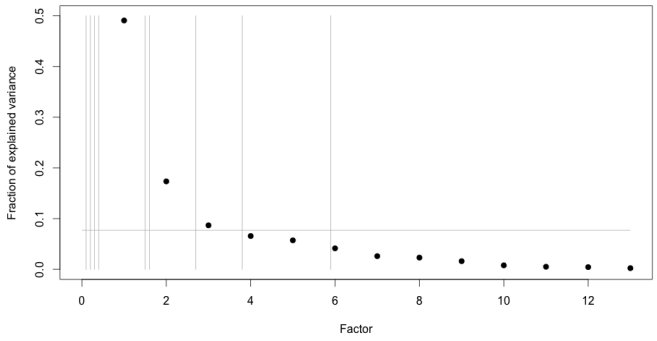

Fig. 1: Screeplot diagram of the variance explained by each factor structure possibility. Vertical lines represent the 10% up to 90% of variance explained by the factor structures.

Correlation patterns were plotted (Figure 2) to demonstrate the relation among the variables in order to help identify the better factor solution. EFA results showed that the four factor structure was able to explain 78% the datasets' variance and was the one with the better factor loading distribution. No variables showed cross-loadings (Figure 3), that is having a loading weight higher than 0.40 in more than one factor, with a difference between loadings less than approximately 0.15. Variables with this behavior tend to refer to more than one latent factor, which might disrupt the specification factor structure of the model. Also, none of the variables showed factor loadings inferior to 0.30, which would indicate that the variables did not relate with any of the latent factors (Figure 3). Only two variables (Hcare3 and EduP1) had factor loadings between 0.30

and 0.50. They were included in the CFAs as there is evidence in the literature indicating that values higher than 0.30 or 0.40 are deemed acceptable [19]. In summary, during the model specification phase (step 1) we noticed that some of the observed variables were not adjusting to latent variables models, which might have influenced the final SEM. Thus from the initial 18 observed variables we ended up with a model constituted by 12 observed variables divided into 4 latent variables.

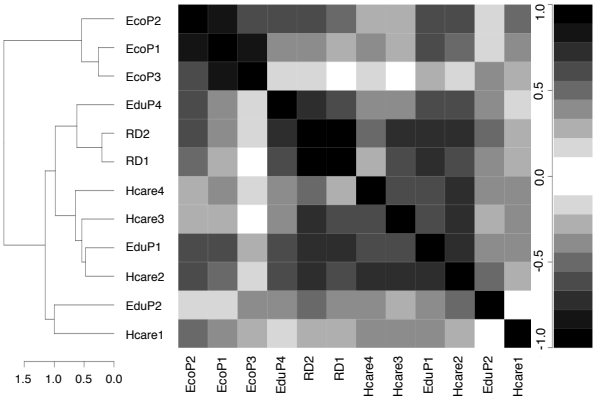

Fig. 2: Correlation matrix for the variables with cluster information on the left. Weight of the coefficients vary from gray (small relation) to black (strong relation).

3.2 Structural Equation Model

Initially we developed a model (Model A, Table 2) with only one exogenous variable (R&D - causing the effect) and three endogenous variables (EduP, EcoP and Hcare - receiving the effect). Exogenous variables are those that originate an effect (path-arrow) to three other variables in the model, while endogenous variables are those that receive the effect (path). However, this model showed poor fitness indicators (Table 2). and several highly correlated residuals. In order to improve this models fit we needed to fix the covariance of errors between variables that would make the model loose its meaning.

Analyzing the residuals behavior directed us to add a secondary path from EduP and EcoP towards Hcare. This is very precise since healthcare performance is known to be influenced by the economic situation of the country and the educational performance [13,26]. Thus, the second model (Model B, Table 2) investigated a direct path from R&D to EduP, EcoP and Hcare, and also a set of paths from EduP and EcoP to Hcare. This model showed problems in its fit indicators (Table 2). In order to improve the models specification and fit indicators, we assessed the modification indices which are indications of the extent of the models fit results that will be improved by adding an additional path to the model. Modification indices suggested the presence of a covariance between the observed variables that constituted R&D and EcoP, therefore affecting the models performance.

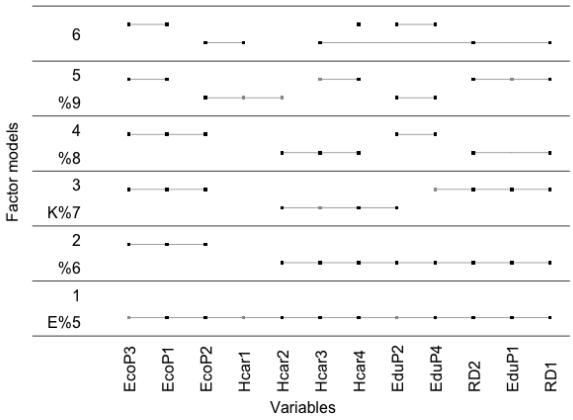

Fig. 3: Factor structures with factor loadings from one up to six factors possibilities. Connected lines indicate possible latent factors.

Then, a third model (Model C, Table 2) was tested (Figure 4), drawing a path between Hcare and the observed variable Edup1 and fixing the covariances errors. This is an interesting relation as it suggests that the healthcare (Hcare) latent construct is also specified by the amount of investment in education (public and private). Model C showed was the one with *best fitness indicators* except for the RMSEA which was above the proposed cutoff point [28,17]. This model was able to explain variance of EduP in 32.5% and HcareP in 49.5%. However, EcoP had only 3.6% of variance explained by the model. The fact that the model was not able to predict the variance of EcoP, suggests that R&D might not be influencing the economic indicators directly as hypothesized. Moreover, there are several methodological problems that may arise in estimating the economic returns to public investment in basic research. According to [26], the main contributions that publicly funded research makes to economic growth are: increasing the stock of useful knowledge; training skilled graduates; creating new scientific instrumentation and methodologies; forming networks and stimulating social interaction; increasing the capacity for scientific and technological problem solving and creating new firms. In [13], the authors state that reviewing the role of education quality in promoting economic growth, conclude that there is strong evidence that cognitive skills are powerfully related to long-run economic growth. Therefore, it is not surprising then to observe a low impact of EcoP. In order to see more important impacts we would need to observe a longer time series.

Path coefficients showed that R&D had moderate to high effect on EduP (0.606) HcareP (-0.511) and small effect on EcoP (0.184). EduP and EcoP also showed small effects on HcareP $(-0.150$ and $-0.056)$. These values might be understood in the same meaning as a regression coefficient (although they are not the same), so the value varies (generally) from 0 to 1, indicating the size of the effect for that specific path. Positive and negative signs indicate the reciprocity of the relation, thus positive values show proportional modulation while negative values indicate inverse relations.

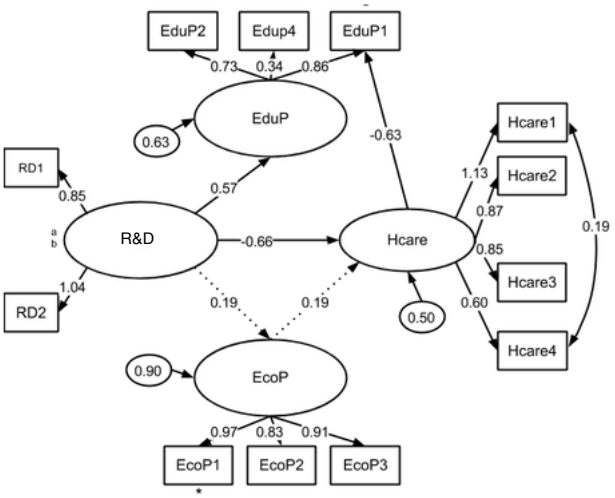

Fig. 4: Structural equation model of the influence of Research and Development on the Economic, Educational and Healthcare performance of EU countries. Values on the arrows connecting latent variables are the path coefficients and indicate the effects weight. Values on the paths connecting observed variables and latent variables are factor loadings Values inside the circles are measurement errors associated with endogenous variables, indicating the extent of variance of that variable that is not explained by the model. Therefore, if the error is high the variable is poorly explained, in this case EcoP.

Finally a fourth model (Model D,Table 2) was tested, mainly for validation purposes, in order to show that our model had a better chance of explaining the relations among the latent variables. This models had EcoP as the main predictor (as an exogenous variable), R&D as a mediator (endogenous and exogenous) and Hcare and EduP as the outcomes. The rationale here is that EcoP is the main predictor of the outcomes (Hcare and EduP) and this effect can be mediated by R&D. However this model did not show a good fitness indicators and the modifications needed to improve its specification could not be accepted because they did not demonstrate theoretical coherence. Therefore, we decided that Model B was the best possible solution to the relations between the latent variables we developed.

After adjusting and modifying the model to find the best possible fit, and also comparing with a different predictive model possibility we verified that R&D positively influences the educational and healthcare systems in the european countries. This results support our initial hypothesized model and suggest that the amount expenditure and personnel in R&D positively influences the educational status in terms of investment, students aid for science/innovation, and also negatively influences the mortality rate, tuberculosis incidence and adolescent fertility. Contradicting our hypothesis, R&D did not directly influence the EU countries' economic performance. However this points to the possibility of other covariants inflicting this model's relations enhancing or impairing the effect of R&D over EU's economic development. Finally it is also noteworthy

that healthcare is also dependent on the investment in education, which perhaps points towards the idea of healthcare coming from educational and social development.

Table 2: CFA Fit Indicators and their respective measurements for all the four models. Model C is adopted in this study.

CFA Fit Indicators	Model A	Model B	Model C*	Model D
CLI	0.771	0.802	0.923	0.835
TLI	0.706	0.768	0.903	0.816
RMSEA	0.218	0.183	0.146	0.188
AIC/BIC	558.73/238.22	529.12/215.05	285.53/-12.40	435.45/118/17

4 Related Work

There have been previous efforts towards utilizing LOD to combine different dataset and analyzing the impact on various factors related to healthcare and research. In particular, the ReDD-Observatory project [30] focused towards integrate several datasets to evaluate the disparity between the amount of research and burden of disease. On the other hand, the Aquameth project [9] provides a new and systematic characterization of 488 European universities The project utilized several micro indicators built on the integrated Aquameth database to characterize the European university landscape according to the following dimensions: history/foundation of university, dynamics of growth, specialization pattern, subject mix, funding composition, offer profile and productivity. Similarly, measurement of the economic impact as well as ranking of universities has been calculated based on a measurement model that relies on several indicators using LOD [25]. However, these studies lacked in terms of comprehensive data to derive output indicators for all economic indicators and policy goals. Moreover, for some of the studies the data was gathered using questionnaires, thus lacking in sufficient amount of data, and was assessed manually.

Additionally, there are studies that have analyzed the relation between funding by the NIH (National Institutes of Health) and burden of disease [11]. In Europe especially, there have been studies to investigate the state of healthcare in the Member States [18]. However, these studies only focused on a particular year as well as gathered data from a limited number of data sources. Moreover, they faced problems such as the availability and quality of data as well as analysis on a limited number of diseases. Also, the data procurement was done manually (online survey and face-to-face discussions with experts) and suffered from incompleteness. Additionally, the Heidi (Health in Europe: Information and Data Interface) data tool[20] presents relevant and comparable information on health at the European level. However, the data is not machine-readable and is also limited to particular years. Similarly, the Pan American Health Organization (PAHO)[21] of the World Health Organization (WHO) launched the EquiLAC project [1]

[20] `http://ec.europa.eu/health/indicators/indicators/index_en.htm`
[21] `http://www.paho.org/`

which measures the equity in health in emphLatin *A*merican and *C*aribbean region in terms of socio-economic disparities, poor utilization of healthcare etc. In contrast, we focus on only European countries spanned over a period of 10 years.

5 Conclusion and Future Work

In this paper, we utilize LOD to evaluate the impact of research and development on economic, educational and healthcare performance particularly in Europe. In particular, we identified two LOD datasets – World Bank and Eurostat; extracted data about relevant variables using SPARQL and fed the retrieved results into a theoretical framework. We identified four particular latent variables to verify the impact: (i) Research and Development (R&D), (ii) Economic Performance (EcoP), (iii) Educational Performance (EduP), (iv) Healthcare performance (HcareP) of the European countries. We used an SEM approach to perform exploratory factor analysis to calculate the correlation between these four latent variables. After performing analysis using different combinations of the variables, we presented a model that showed to have the best goodness of fit. The main objective of our analysis is to show the feasibility and the usefulness of combining different LOD data-sources to assess the impact of R&D by using a structural equation modeling approach. It is beyond the scope of this paper to propose a final model for the assessment of the impact of R&D on the economic, educational and healthcare performance in Europe. Further research and additional data are required to accomplish such an ambitious task. However, the results illustrated the promise of LOD to be utilized in a variety of ways to perform analysis in different domains. Future work will involve application of the model for determining the impact for countries all over the world and over a wider time span. Moreover, we intend to look into ways of streamlining the process of extracting data directly into R (for e.g. via the SPARQL package[22]) and feeding the results automatically into the SEM thus retrieving data from a larger amount of datasets and variables.

References

1. *Investment in Health: Social and Economic Returns*. Pan American Health Organization, 2011.
2. J. D. Adams. Fundamental stocks of knowledge and productivity growth. *Journal of Political Economy*, 4:673–702, 1990.
3. A. assessment of the use of partial least squares structural equation modeling in marketing research. Joe f. hair and marko sarstedt and christian m. ringle and jeannette a. mena. *Journal of the Academy of Marketing Science*, 40(3):414–433, 2012.
4. R. W. Bessette. Measuring the economic impact of university-based research. *Journal of Technology Transfer*, 28:3550361, 2003.
5. K. A. Bollen. *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics, 2000.
6. A. Clarke, M. Gatineau, O. Grimaud, S. Royer-Devaux, N. Wyn-Roberts, I. Le Bis, and G. Lewison. A bibliometric overview of public health research in Europe. *European Journal of Public Health*, 17(1):43–9, 2007.

[22] `http://cran.r-project.org/web/packages/SPARQL/`

7. A. B. Costello and J. W. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research and Evaluation*, 10(7), 2005.
8. P. Cuttance and R. Ecob, editors. *Structural Modeling by example: Applications in Educational, Sociological and Behavioral Research*. Cambridge University Press, 1987.
9. C. Daraio, A. Bonaccorsi, A. Geuna, B. Lepori, L. Bach, P. Bogetoft, M. F. Cardoso, E. Castro-Martinez, G. Crespi, I. F. de Lucio, H. Fried, A. Garcia-Aracil, A. Inzelt, B. Jongbloed, G. Kempkes, P. Llerena, M. Matt, M. Olivares, C. Pohl, T. Raty, M. J. Rosa, C. S. Sarrico, L. Simar, S. Slipersaeter, P. N. Teixeira, and P. V. Eeckaut. The european university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy*, 40(1):148 – 164, 2011.
10. J. Fox. Structural Equation Models. Technical report, R CRAN Packages.
11. L. A. Gillum, C. Gouveia, E. R. Dorsey, M. Pletcher, C. D. Mathers, C. E. McCulloch, and S. C. Johnston. NIH disease funding levels and burden of disease. *PLOS One*, 6(2), 2011.
12. J. Gross. nortest: Tests for normality. Technical report, R CRAN Packages, 2012.
13. E. Hanushek and L. Woessmann. *Education and economic growth*. Elsevier, 2010.
14. R. D. Hays, D. Revicki, and K. S. Coyne. Application of structural equation modeling to health outcomes research. *Eval Health Prof*, 28(3):295–309, 2005.
15. R. Henderson, A. Jaffe, and M. Trajtenberg. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965-1988. *Review of Economics and Statistics*, pages 119 – 127, 1998.
16. K. Hornik. The R project: A language and environment for statistical computing. http://www.r-project.org/, 2008.
17. J. Hox. An intorduction to Structureal Equation Modeling. *Family Science Review*, 11:354–373, 1998.
18. K. Kilpeläinen, A. Tuomi-Nikula, J. Thelen, M. Gissler, A.-P. Sihvonen, P. kramers, and A. Aromaa. Health indicators in europe: availability and data needs. *The European Journal of Public Health*, 2012.
19. P. Kline. *An Easy Guide to Factor Analysis*. Routledge, London, 1994.
20. R. B. Kline. *Principles and Practice of Structural Equation Modeling*. The Guilford Press, New York, 2011.
21. S. Korkmaz. Multivariate normality tests. Technical report, R CRAN Packages, 2013.
22. S. Lab. SCImago institutions rankings. Technical report, SCImago Research Group, 2012.
23. P. O. Larsen and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3):575–603, 2010.
24. E. Mansfield. Academic research underlying industrial innovations: sources, characteristics, and financing. *Review of Economics and Statistics*, 1995.
25. R. Meymandpour and J. G. Davis. Ranking universities using linked open data. In *Linked Data on the Web (LDOW) workshop at WWW*, 2013.
26. A. Salter and B. Martin. The economic benefit of publicly funded basic research: a critical review. *Research Policy*, 2001.
27. J. L. Schafer. Norm package for R, version 3. Technical report, The Methodology Center, The Pennsylvania State University, 2008.
28. K. Schermelleh-Engel, H. Moosbrugger, and H. Muller. Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2):23–74, 2003.
29. T. R. Stats Package. Mahalanobis distance. Technical report, R Documentation.
30. A. Zaveri, R. Pietrobon, S. Auer, J. Lehmann, M. Martin, and T. Ermilov. ReDD-Observatory: Using the web of data for evaluating the research-disease disparity. In *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2011.