INFACT
Joint Action on Health Information

This project is funded by the Health Programme of the European Union

FACT SHEET FOR USE OF ARTIFICIAL INTELLIGENCE FOR PUBLIC HEALTH SURVEILLANCE

WP9 "A case study to develop a Machine Learning (ML)-algorithm to predict the incidence of Diabetes Mellitus and a summary of inspiring examples using linked data and ML techniques"

Key outputs

1. Background: The possibility to link different data sources with each other and the use of artificial intelligence to analyze large datasets are increasing in healthcare. However, linked administrative data can be complex to use and may require advanced expertise and skills in statistical analysis. The main objectives of this study were to develop a generic approach to predict a health outcome from linked data set using machine-learning technique and to identify inspiring examples applying these innovative techniques in public health across European countries.

2. Innovative tools to improve the current EU-HIS: The innovative techniques (i.e., data linkage and/or artificial intelligence) have several advantages, which can improve the current EU-HIS by enhancing completeness and comprehensiveness of information to guide health policy process, by reducing the dimensionality of large datasets and more efficient analysis of large datasets with high precision.

3. Most relevant results: The generic ML-algorithm was a linear discriminant model based on 23 variables related to the biological tests, drugs, medical acts and hospitalization without a procedure over last two years to predict the incidence of diabetes. This algorithm has a sensitivity of 62%, a specificity of 67% and an accuracy of 67% [95%CI: 0.66 – 0.68]. We have identified 16 studies (12 studies related to data linkage, 2 studies applied machine learning and 2 studies used both data linkage and machine learning approaches) as inspiring examples from ten European countries.

4. Feasibility of being integrated in HIS and translated into policies: These results would support countries to learn from each other, to develop, adopt and integrate these innovative approaches to estimating health indicators and to translate those evidence into policy.

Background and Rationale

The possibility to link different data sources with each other and the use of artificial intelligence to analyze large datasets are increasing in healthcare. These innovative techniques (i.e., data linkage and/or artificial intelligence) have several advantages such as data linkage improves completeness and comprehensiveness of information to guide health policy process, whereas the artificial intelligence allows handling data with a large number of dimensions (features) and units (feature vectors) more efficiently with high precision. However, linked administrative data can be complex to use and may require advanced expertise and skills in statistical analysis. The capacity to use data linkage and/or the use of artificial intelligence to estimate and predict health indicators varies across EU-MSs (European Member States). The main objectives of this study were to develop a generic approach to predict a health outcome from linked data set using machine-learning technique and to identify inspiring examples applying these innovative techniques in public health across European countries.

Proposal

To develop the generic approach, we adopted a supervised machine learning approach[1]. Following steps were performed: i. selection of final data set, ii. case/target definition, iii. coding features/variables for a given window of time, iv. split final data into training and test data sets, v. features/variables selection, vi. training model/algorithm, vii. validation of model/algorithm with test data set and viii. selection of the model/algorithm.

The final data set used to develop the ML-algorithm included 44,659 participants and 3468 SNDS variables were coded similarly. Only 23 were selected to train different algorithms. The final algorithm was Linear Discriminant Analysis (LDA) model based on the number of reimbursements of 23 variables related to biological tests, drugs, medical acts and hospitalization without a procedure over last two years to predict the incidence of diabetes. This algorithm has a sensitivity of 62%, a specificity of 67% and an accuracy of 67% [95% CI: 0.66 – 0.68].

We have identified 16 studies (12 studies related to data linkage, 2 studies applied machine learning and 2 studies used both data linkage and machine learning approaches) as inspiring examples from ten European countries[2]. These studies covered 14 different domains of public health. Some of these studies applied classical statistical methods such as multilevel linear regression and some of these studies used artificial intelligence such as machine learning techniques. These studies highlighted that different data collection method, lacking completeness of information or inaccessibility to certain information makes challenging to analysing large linked datasets. Using linked data and AI, the methodological and data analysis aspects can be improved. The results of these studies are used to improve public health surveillance, developing prevention strategies, evaluating health care services and guiding health policy process.

Recommendations for sustainability

We recommend further research to improve the performance of this algorithm to applying on SNDS and to predict the type II diabetes cases in real-time data. More research is needed using various MLTs to predict the incidence of various health conditions by taking into account various determinants (i.e., health and non-health) for improved public health surveillance.

These inspiring examples would support countries to share different experiences and to learn from each other. Furthermore, these examples would help countries to develop, adopt and integrate innovative approaches using data linkage and artificial intelligence to estimating health indicators. These examples also allow comparing various approaches used for innovative use of health information across MSs. These inspiring examples would support to develop the methodological guidelines, which would allow estimating health indicators using linked data and artificial intelligence. Eventually, the evidence produced by using innovative techniques would guide policymaker to make better decisions.

References:

1. *Haneef R., et al. "Use of artificial intelligence for public health surveillance: a case study to develop a machine-learning algorithm to predict the incidence of Diabetes Mellitus", (manuscript in preparation).*
2. *"Inspiring examples from European countries: Implications for innovative use of data sources", InfAct Deliverable 9.2.2, 2020.*