Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210162

ClinFlow - An Interactive Application for Clinical Data Mining

Oana STOICESCU[a,1], Eija FERREIRA[a], Satu TAMMINEN[a], Pekka SIIRTOLA[a], Gunjan CHANDRA[a], Riitta VEIJOLA[b,c] and Juha RÖNING[a]
[a]*Biomimetics and Intelligent Systems Group, Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland*
[b]*PEDEGO Research Unit, Medical Research Centre, Department of Paediatrics, University of Oulu, Oulu, Finland*
[c]*Department of Children and Adolescents, Oulu University Hospital, Oulu, Finland*

Abstract. Analyzing clinical data comes with many challenges. Medical expertise combined with statistical and programming knowledge must go hand-in-hand when applying data mining methods on clinical datasets. This work aims at bridging the gap between clinical expertise and computer science knowledge by providing an application for clinical data analysis with no requirement for statistical programming knowledge. Our tool allows clinical researchers to conduct data processing and visualization in an interactive environment, thus providing an assisting tool for clinical studies. The application was experimentally evaluated with an analysis of Type 1 Diabetes clinical data. The results obtained with the tool are in line with the domain literature, demonstrating the value of our application in data exploration and hypothesis testing.

Keywords. clinical, data, mining, analysis, tool, R, Shiny

1. Introduction

Clinical data is a crucial resource in the process of healthcare progress, fundamental to the development of good care practices for patients and the success of clinical trials [1]. When analyzing clinical data, a thorough understanding of study design particularities combined with medical knowledge can mean the difference between biased and unbiased results [2].

In clinical research, we need in-depth complex analysis methods that are able to identify inconsistencies, as they can have a considerable effect on study results. The amount, distribution, and reasons for inconsistencies may all introduce bias. Machine learning techniques can only be applied after careful hypothesis generation, selection of relevant input cohorts, data preprocessing, feature selection and other fine tuning [3]. To apply these techniques in clinical research, statistics, computer science, and domain knowledge are indispensable. However, these skills rarely coexist. Therefore, complex tools that bridge the gap between domain experts and data scientists are needed [4].

[1] Corresponding Author; Oana Stoicescu, Biomimetics and Intelligent Systems Group, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland; E-mail: oana.stoicescu@oulu.fi.

Some key requirements for successful visualizations of large clinical datasets include dimensionality reduction, interactivity, scalability, fast results and user assistance [5]. A variety of visual analytics tools are currently available for medicine. MAV-Clic [6] and ClinicalVis [7], OpenClinica, openCDMS, TrialDB, and PhOSCo [8] are some of the open-source tools dedicated to management and analytics of medical data. These tools are mainly focused on database management and EHR. While most of them offer visualization and data filtering options, they have limited options for preprocessing, and they do not include dimensionality reduction methods or advanced analytics such as clustering or survival analysis for population statistics.

Our objective is to offer an interactive visual interface that allows the domain expert to process and perform advanced exploratory tasks on clinical trial datasets, without the need for statistical programming knowledge. In this paper, we propose ClinFlow, an open-source prototype built using R [9] and Shiny [10]. To demonstrate the advantages of the application in dataset preparation and rapid hypothesis testing, we perform a small-scale user test using cohort data from the Finnish Type 1 Diabetes Prediction and Prevention (DIPP) study [11], widely studied with the purpose of developing strategies for prevention of Type 1 Diabetes (T1D).

2. Methods

ClinFlow offers a variety of methods, customized for clinical data analyses, which allow the domain expert to build cohorts and to perform exploratory analyses on large clinical datasets. The application includes a module for data processing and filtering, including interactive visual elements for querying the data, checking and deleting erroneous entries, a module for exploratory analysis through a variety of graphical methods, dimensionality reduction, clustering and unsupervised learning including PCA, t-SNE, MDS, SOM, and K-means (This module is refactored from HTPdvis module of the HTPMod application by Dijun Chen [12]. Source: https://github.com/htpmod/HTPmod-shinyApp/), a module for time-series exploratory analysis and panel data creation, and finally, a module for multivariate survival analysis using Cox proportional hazards model. ClinFlow offers interactive graphs that allow users to query the data table directly by hovering, clicking or selecting points in the graphs.

The tool has a modular architecture. Removing, modifying or replacing a module in the code will not affect the rest of the functionalities. Each functionality of the application has its own server module, matched with the corresponding UI module. Figure 1 illustrates an outline of the system architecture.

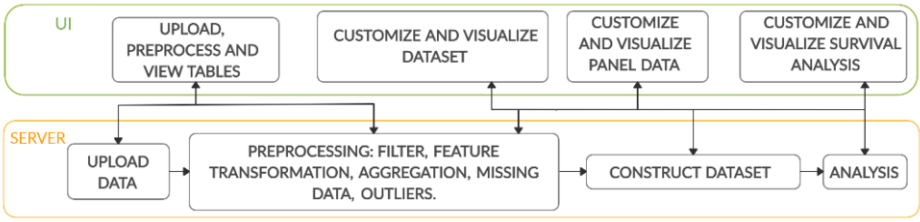

Figure 1. Architecture of ClinFlow.

When data is uploaded into the tool, constant and time-series variables are automatically detected, and data is split into two tables: *patient data* and *visit data*. Visit variables can be summarized over a period of time and added into the *patient data* table - this allows the user to check the impact that values from medical visits from a certain time interval have on an outcome, or compare the values of different groups, even if the subjects had irregular medical visits.

For demonstration, we used ClinFlow to identify the relationships between islet autoantibodies and progression to T1D in the DIPP dataset [11] from Oulu University Hospital. We compared our results with the studies of Knip et al [13] and Pöllanen et al [14], that demonstrate how different types of autoantibodies, autoantibody combinations and age at seroconversion are associated with disease progression.

With the help of the tool, we were able to identify a subset of 253 children with persistent positivity (two or more consecutive positive samples) in at least one autoantibody. 118 of them progressed to T1D before the age of 15 years. The rest were followed up until 15 years old and did not progress to T1D. The subset was created using the data filter, to include children monitored from birth until at least 15 years old, who seroconverted after the age of one year to at least one autoantibody. We excluded the children under 15 years old without a T1D diagnosis and the children who seroconverted before the age of 1 year. The filter option of the tool made it easy to group the data and subset it using the dynamic user inputs (sliders and drop-down lists), and check the summary statistics of the groups. We performed clustering and exploratory analysis to identify the factors associated with the progression to T1D for this subset. We also created a structured panel dataset using the *visit data* by aggregating visit values from certain time intervals to explore the time trends of these values. Additionally, we performed Cox regression survival analysis to visualize the time-to-diagnosis hazard ratios for T1D diagnosis.

3. Results

Using the tool, we were able to identify several associations in the data. We performed PCA and visualized the clusters formed by plotting the principal components (PC). We found a strong relationship between multipositivity (positive samples of two or more autoantibodies occuring at the same time) and progression to T1D, and no relation between single positivity (positive samples of only one autoantibody) and progression to T1D. Figure 2A illustrates the PC values. The points are shaped according to diabetes progression (squares stand for progressors and circles for non-progressors) and the color indicates positivity type (red for multipositivity, yellow for single positivity). The figure suggests an association between multiple seroconversion at an early age, and progression to persistent multipositivity and T1D. Figure 2B shows the relationship between multipositivity (circles) and multiple seroconversion (green). Figure 2C shows a heatmap of loading scores that explain how much each variable contributes to the variation in the data, with seroconversion type contributing the most, followed by the age at seroconversion.

The IAA autoantibody positivity is associated with early seroconversion. In Figure 2D, the density distribution of the seroconversion age shows a peak around the age of 2 years for the groups with IAA at seroconversion. Higher values of IAA autoantibody in early life and high values of GADA later in life are indicative of progression to multipositivity, according to the time trend plots in Figures 2E and 2F.

The survival analysis results in Figure 2G confirm these findings. Additionally, they show an association between multiple seroconversion and an increased risk of rapid progression to T1D. The Kaplan-Meyer curves in Figure 2G show that multiple seroconversion has the lowest time-to-diagnosis.

All of these findings are in line with the results of Knip et al [12] and Pöllanen et al [13], which demonstrates the usefulness of our application in rapidly producing and exporting results from visual analyses, without any manual time-consuming calculations, or thorough statistical processes.

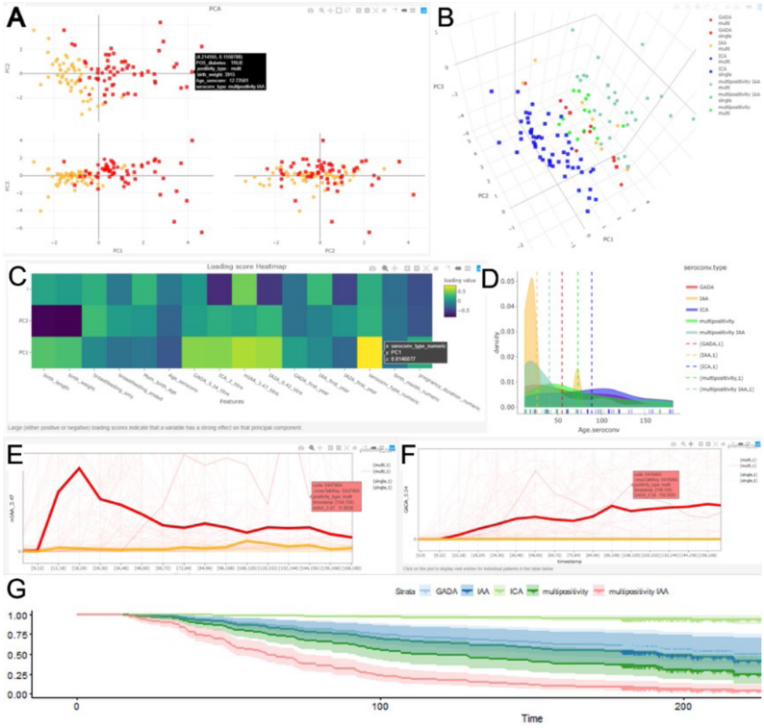

Figure 2. 2D PCA clustering (A). 3D PCA clustering (B). Heatmap of loading scores (C). Density plot of seroconversion age for each autoantibody (D). Time trend of IAA values (E) and GADA values (F) for multipositive (red) and single positivity (yellow). Survival curves, time is the subject's age in months (G).

4. Discussion and Conclusions

The user study performed in this article is an example of how ClinFlow can be used to explore and learn to understand a dataset. With minimal effort and with no programming needed, we were able to identify and prepare the data and test different hypotheses.

The preprocessing module is highly useful in building cohorts and investigating summary statistics for different groups in the data. It allows for visual queries to the data table for checking and deleting anomalies or erroneous entries, as well as mapping the existing information into more interpretable forms, according to user needs.

The visual elements in the tool make it easy to spot associations between groups in the data. The interactive properties of the plots are particularly valuable in allowing the

users to focus on particular details, to zoom in and out of the graphs, to highlight and select points or series in a chart, and to verify the relevant data corresponding to those points, thus providing scalability for larger amounts of data. The reactive properties of this application ensure that the datasets can be explored with only a few clicks, and the visualizations can be adjusted in a matter of seconds. Moreover, this tool allows for cleaning, transformation, in-depth exploration and then exporting of a dataset to various machine learning or deep learning tools for further analysis and verification of initial results found using the visual and interactive capabilities of the tool.

ClinFlow seeks to fill the gap between information technology and clinical research. It provides an interactive interface that allows the clinical researcher to include domain expertise into the analysis process, without the need for statistical programming knowledge. Open-source code of this project can be found at: http://bit.ly/3uZFint along with a list of the R packages used to build this tool.

A comprehensive usability study is planned for the future, where clinical researchers will be recruited to perform a multitude of tasks with ClinFlow, and quantitative measures on the use of this tool will be collected.

References

[1] Institute of Medicine (US) Roundtable on Value \and Science-Driven Health Care. Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary. Washington (DC): National Academies Press (US); 2010.
[2] Herbert RD, Kasza J, Bø K. Analysis of randomised trials with long-term follow-up. BMC Med Res Methodol. 2018;18(1):48.
[3] Ljubic B, Roychoudhury S, Hang Cao X, Pavlovski M, Obradovic S, Nair R, Glass L, Obradovic Z. Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. Comput Meth Prog Bio.2020;197.
[4] Rokošná J, Babič F, Majnarić LT, Pusztová L. Cooperation Between Data Analysts and Medical Experts: A Case Study. In: Holzinger A, Kieseberg P, Tjoa A, Weippl E (eds) Machine Learning and Knowledge Extraction. LNCS, 2020;12279. Springer, Cham.
[5] Lowe J, Matthee M. Requirements of Data Visualisation Tools to Analyse Big Data: A Structured Literature Review. In: Hattingh M et al (eds). Responsible Design, Implementation and Use of Information and Communication Technology. LNCS, 2020;12066. Springer, Cham.
[6] Zeeshan A, Minjung K, Bruce TL. MAV-clic: management, analysis, and visualization of clinical data. JAMIA 2019;2(1):23–28.
[7] Ghassemi, M., Pushkarna, M., Wexler, J., Johnson, J., Varghese, P. ClinicalVis: Supporting Clinical Task-Focused Design Evaluation. ARXIV. 2018.
[8] Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS. Data management in clinical research: An overview. Indian J Pharmacol. 2012;44(2):168-72.
[9] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2020. Available at http://www.R-project.org/.
[10] Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R. R package version 1.5.0, 2020. Available at http://shiny.rstudio.com/.
[11] Haller, MJ, Schatz, DA. The DIPP project: 20 years of discovery in type 1 diabetes. Pediatr Diabetes, 2016;17: 5-7.
[12] Chen D, Fu LY, Hu D, Klukas C, Chen M, Kaufmann K. TheHTPmod Shiny application enables modeling and visualization of large-scalebiological data. Communications Biology 2018;1.
[13] Knip M, Siljander H, Ilonen J, Simell O and Veijola R. Role of humoral beta‐cell autoimmunity in type 1 diabetes. Pediatr Diabetes, 2016;17: 17-24.
[14] Pöllänen PM, Lempainen J, Laine AP, Toppari J, Veijola R, Vähäsalo P, Ilonen J, Siljander H, Knip M. Characterisation of rapid progressors to type 1 diabetes among children with HLA-conferred disease susceptibility. Diabetologia. 2017;60(7):1284-1293.