A Large-scale and Extensible Platform for Precision Medicine Research

Fodil Belghait
École de Technologie Supérieure
1100 Notre-Dame West
Montréal, Canada
fodil.belghait.1@ens.etsmtl.ca

Alain April
École de Technologie Supérieure
1100 Notre-Dame West
Montréal, Canada
alain.april@etsmtl.ca

Pavel Hamet
Centre hospitalier de l'Université de Montréal
Montréal, Canada
pavel.hamet@umontreal.ca

Johanne Tremblay
Centre hospitalier de l'Université de Montréal
Montréal, Canada
johanne.tremblay@umontreal.ca

Christian Desrosiers
École de Technologie Supérieure
1100 Notre-Dame West
Montréal, Canada
christian.desrosiers@etsmtl.ca

ABSTRACT

The massive adoption of high-throughput genomics, deep sequencing technologies and big data technologies have made possible the era of precision medicine. However, the volume of data and its complexity remain important challenges for precision medicine research, hindering development in this field. The literature on precision medicine research describes a few platforms to support specific types of studies, but none of these offer researchers the level of customization required to meet their specific needs [1].

Methods: We propose to design and develop a platform able to import and integrate a very large volume of genetics, clinical, demographical and environmental data in a cloud computing infrastructure. In our previous publication, we presented an approach that can customize existing data models to fit any precision medicine research data requirement [1] and the requirement for future large-scale precision medicine platforms, in terms of data extensibility and the scalability of processing on demand. We also proposed a framework to meet the specific requirement of any precision medicine research [2]. In this paper, we describe how this new framework was implemented and trialed by the precision medicine researchers at the Centre Hospitalier Universitaire de l'Université de Montréal (CHUM).

Results: The data analysis simulations showed that the random forest algorithm presents better accuracy results. We obtained an F1-Score of 72% for random forest, 69% using linear regression and 62% using the neural network algorithm.

Conclusion: The results suggest that the proposed precision medicine data analysis platform allows researchers to configure, prepare the analysis environment and customize the platform data model to their specific research in very optimal delays, at very low cost and with minimal technical skills.

DPH' 19, November 20–23, 2019, Marseille, France

ACM ISBN 978-1-4503-7208-4/19/11…$15.00
https://doi.org/10.1145/3357729.3357742

CCS CONCEPTS

• Database Theory → Database structures and algorithms; Data Extraction Transform and Load

KEYWORDS

Clinical Databases, Genomics, Precision medicine, Bioinformatics, Big Data

ACM Reference format:

Fodil Belghait, Alain April, Pavel Hamet, Johanne Tremblay, Christian Desrosiers. 2019. A Large-scale and extensible platform for precision medicine research. In Proceedings of 9th International Digital Public Health Conference (DPH'19), November 20–23, 2019, Marseille, France. ACM, New York, NY, USA. https://doi.org/10.1145/3357729.3357742

1. Introduction

The main goal of biomedical research and medical practice is the improvement of human health. From these two disciplines emerged a novel research field named precision medicine. Precision medicine aims at creating custom treatments for individuals based on their personal genetic, biomarker, phenotypic, psychosocial and lifestyle characteristics [3]. The

capacity to collect this data increased exponentially with the massive adoption of high-throughput genomics, deep sequencing technologies and big data technologies. However, the field of precision medicine is lacking in organizing and cheaply processing these increasing volumes of data so as to be used commonly with patients [4].

This paper aims at investigating how a new proposed data analysis framework can be used in the context of a precision medicine research case study in order to evaluate its potential. An actual case study will be made using the platform presented in [2].

2. Background

Lately, several precision medicine platforms have emerged that propose solutions for researchers to support them in collecting, managing and analyzing genomic and clinical data for their research. Table 1 summarizes the current landscape where we have highlighted a number of desired characteristics for comparison. The broadness of the data used by precision medicine research involves an extremely large variety of data elements. The technological challenges of capturing, structuring and processing this data create bottlenecks, impeding the implementation of precision medicine research for many organizations and slowing down its use in clinical care [5].

Using published papers about these precision medicine research platforms, we have reviewed the features of each platform and assessed them for seven key features that are, from our perspective, a prerequisite to a desired platform. The objective of this analysis is to determine whether they support the proposed platform targeted features [2].

Table 1 Precision medicine platform landscape February 2019

Precision Medicine Platforms	BRISK[6]	iCOD[7]	I2B2/TranSMART[8]	OncDRS[10] [11]	deCODE[12]	IROmiCS[13], [14]
Allows the use of Omics Data	Y	Y	Y	Y	Y	Y
Allows the use of other patient/env. data	Y	Y	Y	Y	Y	Y
Allows its data model to adjusted/extended	N	N	N	N	N	N
Provides a flexible Infrastructure	N	N	N	N	N	N
Allows study on large amount of data	Y	Y	Y	Y	Y	Y
Allow computing Infrastructure to scale	N	N	N	N	N	N
Has a function to reproduce a study	Y	Y	Y	Y	Y	Y

Our analysis showed that none of the reviewed platforms supports the data model extension, platform flexibility or scalability as a native function. In addition to these missing features, we have identified two technological challenges that have to be overcome concerning the data processed by these platforms: first, to be able to manage the data complexity/heterogeneity; and second, to ensure that this data can be integrated in a single location.

3. New Proposed Platform for Precision Medicine Research

The new platform has been designed and developed to offer all the features analyzed in the Table 1 above and attain the following four objectives:

1. **Flexibility of the data schema:** The data schema should be customized easily and quickly to fit any precision medicine goals. Also, the platform software architecture should be loosely coupled to the cloud computing provider technology to allow for interoperability (i.e. it should easily work with the technology of other cloud providers).
2. **Scalability**: It is required in three main areas: first, the ability to import large amounts of data–the large amount of patient data (i.e. clinical, genetic) needs to be easily migrated/loaded into a customized data model within a reasonable amount of time and cost; second, IT infrastructure scalability–the IT infrastructure has to be easily scalable to handle large volumes of data, i.e. adding and removing computing instances should be easy and turning the infrastructure on/off should allow cost savings; and third, the research data analysis scalability–the framework aims to allow researchers to reuse their existing data analysis and scale it with new data analysis requirements.
3. **Usability:** The proposed framework APIs have to be simple and easily used by individuals that are not IT specialists; this means its operation should not require advanced database and computer infrastructure skills.
4. **Reproducibility:** The framework needs to allow the researcher to easily replicate the result of any prior data analysis using the same data samples and analysis requirements.

The new platform has been designed with three main layers. The first is to collect and build a customized research data model, the second is to migrate and integrate the collected data into the new data model and the last is for the data analysis and knowledge extraction [2]. To use the new platform, we have proposed a new pipeline for the future precision medicine data analysis based on eight steps: 1) research data analysis requirement definition, 2) customized research data model creation, 3) data migration cluster setup, 4) data migration, 5) data migration cluster scaling, 6) data analysis cluster setup, 7) data analysis and 8) data analysis cluster scaling [2].

The proposed pipeline and platform enabled us to execute all of the case study data preparation and analysis steps within a very short time and with low cost. The platform configuration and data preparation process for 71 MB of clinical data and 101.578 GB of genetics data using m4.4xlarge instances of Amazon cloud computing took approximately 3.5 hours: 2 minutes for data module customization, 8 minutes for the platform configuration, 1.4 minutes for clinical data migration and 1 hour and 18 minutes for genetics data migration. The cost of the whole process,

including the transactions and data storage and the research execution cost, was under $40 USD [2].

4. The case study

The case study proposes using the new platform process and tools to conduct a precision medicine data analysis. It involved the creation of a predictive model for the risk of development and the progression of renal complications in patients with Type 2 Diabetes (T2D) using the patients' genomic variants, clinical and environmental data with Neural Network (NN), random forest (RF) and logistic regression (LR) algorithms. It uses a list of informative genetic variants encompassing relevant risk factors for renal complications, selected from publicly available GWAS data and tests them on the ADVANCE cohort data [15], [16]. The study was performed following the steps of the proposed platform pipeline [2].

4.1 Data and research analysis–understanding and execution

The steps involved in understanding and preparing the data require an important effort in these types of studies. In this case study, the ADVANCE dataset was originally provided to us in flat files: 1) a single flat file in a fixed-width text format containing the patients' identification number and gender; 2) 70 flat files, using a variable-width text format. As shown in Figure 1 below, each row contained the genotype data for a single nucleotide polymorphism (SNP). The first five entries of each line are: 1) the SNP ID; 2) the RS number of the SNP; 3) the base-pair position of the SNP; 4) the allele coded A; and 5) the allele coded B. The next three numbers on the line represented the probabilities of the three genotypes, AA, AB, and BB, at the SNP for the first individual in the cohort. The following three numbers are the genotype probabilities for the second individual in the cohort. The next three numbers represent the third individual and so on. The order of individuals in the genotype file matched the order of the individuals in the sample file. Also, the probabilities need not to add up to 1, to allow for the possibility of a NULL genotype call. This format allowed for genotype uncertainty. Additionally, some data was available in a Postgres [17] database containing all the clinical and socio-demographic data of the 4098 genotyped patients.

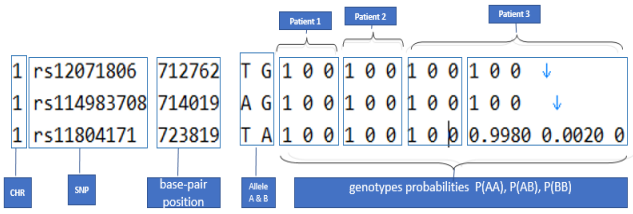

Figure 1 - Genotype File Format

4.2 Identifying risk factors and associated SNP's

Albuminuria and a reduced estimated glomerular filtration rate (eGFR) are manifestations of diabetic nephropathy that predict end-stage renal disease requiring dialysis or transplantation. Based on the literature review, the researcher identified 76 SNPs associated to low eGFR levels or decline and albuminuria (UACR) [15],[17] two markers of renal damage in building the predictive model.

4.3 Clustering patients in data sample

Kidney diseases are commonly found in diabetic people; 50% of diabetic patients show signs of renal damage in their lifetime [16],[15]. It is the leading cause of chronic kidney disease (CKD) in Canada [15]. CKD can be seen in a variety of conditions, including diabetes and high blood pressure. Measuring glomerular filtration rate is considered the most accurate way to detect changes in kidney condition. The eGFR is a calculation based on a serum creatinine test. Creatinine is a muscle waste product that is filtered from the blood by the kidneys and released into urine at a relatively steady rate. When kidney function decreases, less creatinine is eliminated, and its concentration in blood increases. With the creatinine test, a reasonable estimate of the actual eGFR can be obtained. The patients will be clustered based on the eGFR decline value, with a value of 60 ml/min/1.73m2 or higher comprising the normal range. Any patient with eGFR = 60 or below for at least three months (three measures in the clinical database) will be identified as a CKD case.

4.4 Data pre-processing, transformation and reduction

This activity aims to ensure that the data quality is suitable for running artificial intelligence (e.g. machine learning) algorithms. First, the data collected had to be migrated out of the clinical relational database (e.g. the Postgres database) and the many genotype flat files had to be transformed and integrated into the extended ADAM [19] schema. This meant removing all the erroneous data, converting Boolean to binary data format, preparing the data so that it can be used by the three prediction algorithms (i.e. normalize the age, gender and region columns), and reduce empty and unnecessary data analysis columns. At the end of this step, all this data was converted into a single table comprising of 1118 rows (1 row per patient) and 112 columns, with 76 columns used for genetic data and 36 columns for clinical data (i.e. age, gender, geo-ethnic region).

This table was created using three datasets: 1) clinical data set (2394 patients), 2) genetics data set with over 15 billion rows and 3) the SNP list associated with the eGFR risk group comprising of 5 columns and 76 rows. The table contained two calculated fields. The risk allele genotype (RAG) refers to the values calculated for each SNP that assess the SNP's impact on the disease. They have been calculated (see Algorithm 1 below) using the probabilities of the reference and alternate allele along with the risk allele column for each SNP.

Algorithm 1:
If the risk allele = reference allele (RA)
then RiskAlleleGenotype = 2 * P(RA, RA) + P(RA, AA)
else RiskAlleleGenotype = 2 * P(AA, AA) + P(RA, AA)

Where P(RA,RA), P(RA,AA), and P(AA,AA) represent the

The diagnosis results in a Boolean value that will be used for the classification label. It has been calculated (see Algorithm 2 below) using the eGFR measure from the clinical database.

Algorithm 2:
If eGFR ≤ 60ml/min/1.73m2
then CKD = 1
else CKD = 0

4.5 Prediction models used in analysis

In the analysis stage of this case study, we experimented with three classification models. We selected these models because of their popularity in recently published literature pertaining to gene selection [20],[21][22].

1. Random Forests (RF): This model for regression and classification was introduced for the first time in 2001.Since then, it has gained tremendous popularity and has become one of the most commonly used classification approaches, competing with logistic regression in this regard. It is commonly used in bioinformatics because of its prediction accuracy and model interpretability [58]. It is a tree-based machine learning algorithm, well adapted to problems with few data points and many features [59].
2. **Neural Network (NN):** This model consists in feed-forward multilayer network [60] that attempts to emulate the biological
3. network of neurons in the brain by connecting the predictor and the response variables using layers of intermediate (hidden) nodes. In our analysis, we used two input layers. The input layer represents the features used in the analysis. The features used for the analysis are: the RS number of the SNPs, the gender, the age and to make the patient's diagnosis we used the data of the CKD column (CKD = 1, Non-CKD = 0).
4. **Logistic Regression (LR):** This model is a generalization of linear regression [61]. It is used mainly to predict binary or multi-class dependent variables. It requires the response variable to be discrete. It is widely used in biostatistical applications in which binary responses (two classes) occur quite frequently. For example, patients survive or die, have heart disease or not, or a condition is present or absent.

4.6 eGFR decline prediction model results

The models were executed and evaluated based on the accuracy measures. The results of Table 2 were achieved using an average of 10 simulations for each model. The data sample of 1118 patients was divided into two groups: 75% for training the models and 25% for testing them. Even with a relatively small sample size, the predictive models demonstrated better results with the Random Forest algorithm than with Linear Regression (LR) or Neural Network (NN). We found that the NN model achieved a classification accuracy of 56,15% with a precision of 65,70%, recall of 59,28%, and an F1-score of 62,28%. LR produced better results than the NN, achieving a classification accuracy of 62,54% with a precision of 69,71%, recall of 68,54% and an F1-score of 68,12%. However, the RF model achieved a classification accuracy of 64,60% with a precision value of 68,66%, recall of 77,53%, and an F1-score of 72,82%. Table 2 presents the complete set of results in a tabular format.

Table 2 Tabular results of the training of the predictive models

	Neural Network (NN)				Linear Regression (LR)				Random Forest (RF)			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
1	0,5430	0,6380	0,5843	0,6100	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
2	0,5739	0,6688	0,6011	0,6331	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
3	0,5704	0,6626	0,6067	0,6334	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
4	0,5326	0,6364	0,5506	0,5904	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
5	0,5704	0,6646	0,6011	0,6313	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
6	0,5533	0,6500	0,5843	0,6154	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
7	0,5670	0,6605	0,6011	0,6294	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
8	0,5704	0,6626	0,6067	0,6334	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
9	0,5567	0,6561	0,5787	0,6149	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
10	0,5773	0,6708	0,6067	0,6372	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
Avg	0,5615	0,6570	0,5921	0,6228	0,6254	0,6971	0,6854	0,6912	0,6460	0,6866	0,7753	0,7282
accuracy = (tp + tn) / (tp + tn + fp + fn); precision = tp / (tp + fp); recall = tp / (tp + fn); f1_score = 2 * precision * recall / (precision + recall)												

5. Case study results and discussion

The purpose of this case study was to demonstrate the feasibility of the proposed approach and validate the expected benefits of using the proposed framework with a tangible precision medicine study. We wanted to assess its usability, flexibility, scalability, and the reproducibility of the study. The study also demonstrated how each design characteristic that was implemented and tested earlier this year could perform in a real precision medicine study. The case study was conducted to try to answer the following questions about whether the proposed framework allows: 1) the extensibility and scalability of the data model and data storage; 2) the scalability of the data migration and analysis infrastructure to handle large amounts of data effectively; 3) what technical skill level is required from researchers to use this precision medicine framework; and 4) can the analysis be repeated easily using this proposed framework.

Table 3 - Case study results

Analysis step	Type Manual /Automated	Time	Portability to other cloud	Case study feasibility	Comment
PM research data analysis requirement definition	M		N/A	N/A	
Extend the current genetic data model to add case study data requirement (two additional data schema)	A	2 min	Y	Y	
Data migration cluster setup	A	8 minutes	Y	Y	
Data migration and integration	A	3 hours 18 minutes	Y	Y	This time was required to migrate and integrate the clinical and genetic data using a cluster of 10 instances (m4.4xlarge) and data of 1118 patients (101 GB)
Data Analysis cluster setup	A		No	Y	Customization required to configuration files to allow the portability of this API to other cloud computing providers
Scaling data analysis framework cluster	A	50 secs to add 1 instance, 3 min to add 5 instances	No		The Flintrock tool to manage the spark instances, it runs only on AWS cloud.

5.1 Case study results overview

Table 3 lists the results obtained for each of the framework steps of Figure 1.

5.2 Research Data Analysis process evaluation

The motivation of this work was to create a flexible and scalable framework for precision medicine that allows researchers to repeat data analyses from previous studies with ease. The case study demonstrates that the combination of big data technologies and cloud computing can offer an extremely scalable and adaptable framework for data scientists and researchers in precision medicine, providing the following benefits:

1. **Framework data model flexibility:** Once the list of use case data requirements has been identified, we have automatically extended the latest version of the data model ofADAM Genomics with the new, defined data model using the data model extension "API." The extension process took less than 10 minutes. An important portion of this time was spent adding the new data fields into the data model definition file. The flexibility of the process goes beyond the extension of ADAM schema and allows for the extension of any existing Avro schema with new data analysis requirements, which will allow researchers to follow an incremental process in their precision medicine data analysis.
2. **Framework infrastructure flexibility:** We used the Amazon AWS cloud infrastructure to fulfill the use case data analysis; however, the proposed frameworks APIs are independent from the AWS custom services. We have used new Linux instances that only have the OS installed and can be provided by any cloud computing infrastructure provider. The S3 archive service has been employed to store the framework data. If, for any reason, the researcher decides to use another cloud data storage solution, he or she can do so with few modifications to the APIs.
3. **Framework data scalability:** Two separate APIs were used to migrate the clinical and genetics data of 1118 patients into

the new data model. The APIs are designed to always add new data into the data model, which allows researchers to scale their data samples with any new data they may acquire in the future, and run their analysis against the new data samples.

4. Framework infrastructure scalability: The configuration of the use case data migration and analysis infrastructure using the PROPOSED FRAMEWORK APIs was completed in a very short period of time. The cluster setup and configuration required less than 10 minutes for 5 instances. The time required for data analysis infrastructure cluster management (adding and removing instances from the cluster) varied from 50 seconds for adding a single instance to 3 minutes for adding 5 instances to the cluster. The framework offers tremendous flexibility to researchers in customizing the size of the cluster in real-time, depending on the resources required by the current data analysis steps.
5. **Research data analysis reproducibility:** The use case data analysis can be replicated at any time and only requires the initiation of new data analysis clusters using the framework APIs. All the migrated data is stored in S3 buckets and can be used as long as it is stored in the bucket. In addition, all the temporary data is persisted in S3, which will help the researcher improve the performance of his data analysis process owing to the fact he is not required to recreate them for every simulation.

The case study demonstrated that, using the PROPOSED FRAMEWORK, the total cost was around 8$ USD. For the data migration, we employed a cluster of 10 m4.4xlarge Linux instances (one master node and nine workers). The cluster allocated 144 cores and 557 GB of memory in total in order to process the data conversion and migration into the framework data model. Using this big data infrastructure, the process of converting and migrating clinical data took 1.4 minutes; however, processing genetics data took 3.3 hours. The cost of the data transformation and migration was around 8$ USD (Time = 4 hours including the configuration time, 0.20$ USD per hour for each instance, for a total of 10 instances).We have not evaluated the cost of the data analysis but, as in the case of the migration, without combining the two technologies, each analysis would be extremely difficult and expensive. Most of the current precision medicine platforms use their own big data infrastructure, which limits the scalability of their platform and prevents them from processing larger volumes of data. The PROPOSED FRAMEWORK is the opposite of this, offering researchers an optimal platform with no limit in terms of scalability and portability to any cloud technology owing to its infrastructure flexibility (loosely coupled to specific cloud computing technology).

5.3 Case study data analysis–feasibility with other precision medicine frameworks

Table 4 shows the feasibility of the proof of concept analysis compared with the best-known precision medicine and bioinformatics frameworks. The objective was to evaluate the feasibility of the case study with these frameworks based on the 6 design characteristics we have targeted for our framework. Table 4 Case study feasibility with precision medicine framework review Precision Medicine Software Precision Medicine platform features New Proposed Platform

Table 4 Case study feasibility with precision medicine framework review

Precision Medicine Software / Precision Medicine platform features	New Proposed Platform	BRISK[6]	iCOD[7]	I2B2/TranSMART[8] [9]	OncDRS[10] [11]	deCODE[12]	IROmiCS[13], [14]
Omics Data	Y	Y	Y	Y	Y	Y	Y
Non-genomic PM data	Y	Y	Y	Y	Y	Y	Y
Data model extensibility	Y	N	N	N	N	N	N
Infrastructure flexibility	Y	N	N	N	N	N	N
Data scalability	Y	Y	Y	Y	Y	Y	Y
Infrastructure scalability	Y	N	N	N	N	N	N
Research Reproducibility	Y	N	N	N	N	N	N
Feasibility of the proof of concept PM data analysis	Y	N	N	N	N	N	N

The case study feasibility analysis (Table 4) showed that the proposed framework is the only one capable of fulfilling the proposed analysis without any change, and it is scalable for any other type of precision medicine research data analysis, whereas all of the existing frameworks reviewed fail in at least one design

objective that we have targeted in our research. From the reviewed frameworks listed, no one supports the dynamic flexibility of their data model for new data requirements. On the other hand, only 2 are implemented using big data with cloud infrastructure to support the scalability required to process large scale data, but they are tightly coupled to the cloud computing technology used.

6. Conclusion

The purpose of this proposed precision medicine framework is to mitigate the technical challenges involved in precision medicine research and allow the large-scale data analysis process to be as simple as possible for the researchers. This framework could allow researchers to better focus on the analysis of the patient data rather than on the many technical issues. In this paper, we presented a promising framework comprising processes and toolsets to be used in any large-scale precision medicine research. We conducted an initial validation of the proposal with a real experimental case study that needed to adapt the data analysis schema and load a large quantity of data, ready for large-scale analysis. This first validation confirms four main advantages of the proposed approach: 1) the possibility to adapt and use the latest version of the ADAM schema using any additional data requirement that a particular precision medicine study requires; 2) the simplification of the many technical tasks involved in the data analysis process by using the four processes: create, setup, scale the data migration and conduct the data analysis without too much technical knowledge; 3) the ease of repeatability of this study: by ensuring that all the intermediate and final data as well as results are stored in an S3 location that ensures the repeatability of the data analysis without the need for recalculation of intermediate results; and 4) the portability of the framework to any cloud computing infrastructure. In the case study, we used AWS services; however, we could have employed any other cloud infrastructure; the framework's automated processes do not exploit any custom services of Amazon beside S3 and AIM services. This objective of the study was partially achieved. In order to make the platform APIs work on another cloud computing platform, we will need to adapt the APIs to the new cloud computing service provider for the data storage and for identity access management.

ACKNOWLEDGMENTS

This research project was conducted without government or industry funding. The results of this precision medicine case study were independently verified by the bioinformatics staff investigating diabetes in Dr. Pavel Ahmet's research lab located within the Centre de recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM). They also provided us with the data required to conduct this case study (i.e. determine a predictor for an early diagnosis low eGFR Renal impairment in patients with T2D).

REFERENCES

[1] F. Belghait, B. Kanzki, and A. April, "ADAM Genomics Schema - extension for precision medicine research *," in *DH '18 Proceedings of the 2018 International Conference on Digital Health*, 2018, p. 4.

[2] F. Belghait and A. April, "The Future of Large-Scale Precision Medicine Research Platforms : Preparing the Data for Analysis," International Journal of trends in Research and Development (IJTRD), issn:2394-9333, Special Issue, December 2018: 91–94.

[3] N. Institutes, O. F. Health, and P. Medicine, "DIABETES CARE SYMPOSIUM NIH Precision Medicine Initiative : Implications for Diabetes Research," pp. 1–5, 2016.

[4] A. A. Morgan, D. C. Crawford, J. C. Denny, and S. D. Mooney, "PRECISION MEDICINE: DATA AND DISCOVERY FOR IMPROVED HEALTH AND THERAPY," pp. 1–6, 2018.

[5] O. Wolkenhauer *et al.*, "Enabling multiscale modeling in systems medicine," *Genome Med.*, vol. 6, no. 3, pp. 1–3, 2014.

[6] A. Tan, B. Tripp, and D. Daley, "BRISK-research-oriented storage kit for biology-related data," *Bioinformatics*, vol. 27, no. 17, pp. 2422–2425, 2011.

[7] K. Shimokawa *et al.*, "ICOD: An integrated clinical omics database based on the systems-pathology view of disease," *BMC Genomics*, vol. 11, no. SUPPL. 4, p. S19, 2010.

[8] E. Scheufele *et al.*, "tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform.," *AMIA Jt Summits Transl Sci Proc*, vol. 2014, pp. 96–101, 2014.

[9] S. N. Murphy, M. E. Mendis, D. A. Berkowitz, I. Kohane, and H. C. Chueh, "Integration of clinical and genetic data in the i2b2 architecture," *AMIA Annu Symp Proc*, no. 2, p. 1040, 2006.

[10] J. Orechia *et al.*, "Applied & Translational Genomics OncDRS : An integrative clinical and genomic data platform for enabling translational research and precision medicine," *ATG*, vol. 6, pp. 18–25, 2015.

[11] E. R. Londin and C. I. Barash, "Applied & Translational Genomics What is translational bioinformatics ?," *ATG*, vol. 6, pp. 1–2, 2015.

[12] H. Hakonarson, J. J. R. Gulcher, and K. Stefansson, "deCODE genetics, Inc.," *Pharmacogenomics*, vol. 4, no. 2, pp. 209–15, 2003.

[13] L. F. Soualmia, J. Darmoni, and L. F. Soualmia, "recherche d ' information dans le Dossier Patient e To cite this version : Intégration de données cliniques et omiques pour la recherche d ' information dans le Dossier Patient Informatisé," 2015.

[14] S. J. D. Chloé Cabot, Lina F. Soualmia, "Intégration de données cliniques et omiques pour la recherche d'information dans le Dossier Patient Informatisé," in *Collection AFIA. Journées Francophones d'Ingénierie des Connaissances - IC 2015, Jul 2015*, 2015.

[15] S. R. Heller, "A summary of the ADVANCE Trial.,"

Diabetes Care, vol. 32 Suppl 2, pp. 1–5, 2009.

[16] ADVANCE Management Committee, "Study rationale and design of the ADVANCE study: a randomised trial of blood pressure lowering and intensive glucose control in high-risk individuals with type 2 diabetes mellitus. Action in Diabetes and Vascular Disease: Preterax and Diamicron Modified-R," *Diabetologia*, vol. 44, pp. 1118–1120, 2001.

[17] A. Patel, J. Chalmers, and N. Poulter, "ADVANCE: Action in diabetes and vascular disease," *J. Hum. Hypertens.*, vol. 19, pp. S27–S32, 2005.

[18] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, pp. 1–13, 2006.

[19] J. G. Liao and K.-V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case," *Bioinformatics*, vol. 23, no. 15, pp. 1945–1951, 2007.

[20] A. Endo, T. Shibata, and H. Tanaka, "Comparison of Seven Algorithms to Predict Breast Cancer Survival," vol. 13, no. 2, pp. 11–16, 2008.