Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

The Future of Large-Scale Precision Medicine Research Platforms: Preparing the Data for Analysis

[1]Fodil Belghait and [2]Alain April,
[1,2]École de Technologie Supérieure, Montréal, CANADA

Abstract: The extensive adoption of high-throughput genomics, microarray, and deep sequencing technologies has accelerated the possibility of more complex precision medicine research using very large amounts of heterogeneous data [1]. The availability of this data allows data scientists and clinicians to develop tailored individual strategies. Therapeutic and preventive treatments can be proposed, with greater accuracy, targeting subgroups of patients for specific illnesses using large amounts of genomic, clinical, lifestyle, and environment data [2]. Next generation sequencing (NGS) technology is key in supporting precision medicine research; however, the data's volume and complexity poses challenges for its clinical application [3]. While Big Data's analytics could uncover hidden patterns, new correlations, and other insights through the examination of large-scale data sets, it is still difficult to master [4]. In this paper, we present what is required of future large-scale precision medicine platforms in terms of data extensibility and the scalability of processing on demand. It presents a proposed platform architecture as well as open-source Big Data technologies that would allow to easily enrich a flexible data schema, provide the power needed to load large amounts of data and make this centralized database available for specific precision medicine research.

Keywords: *Precision medicine; Genotyping; Clinical database; Cloud computing; Big Data; Bioinformatics.*

I. INTRODUCTION

The terms precision or personalized medicine, translational medicine and translational research are used interchangeably in the literature [5][6][7][8]. In this paper, we use the term precision medicine to refer to an emerging research field aimed at personalized disease treatment and prevention that takes into account individual differences in genes, patient treatment history, environment, and lifestyle. It integrates the advancements in molecular biology using clinical trials [9][10]. This personalized approach to patient treatments will allow doctors and researchers to predict, with greater accuracy, the course of treatment and prevention strategies to apply for a particular disease for specific groups of individuals. It will also allow for the creation of therapeutic and preventive strategies targeting the specific needs of patients on the basis of genetic, biomarker, phenotypic and psychosocial characteristics, which distinguish one patient from another [2]. For precision medicine research to be effective, computational models that integrate data and knowledge from both clinical and genetic research in order to gain a better understanding of disease are required [11].

In recent years, numerous precision medicine platforms have emerged that propose innovative solutions to collect, manage, and analyse large amounts of genomic and clinical data to be used in precision medicine research. These require that the researcher has access to electronic healthcare records (EHR) that contain patient clinical data. These platforms often offer functionality and programming frameworks that are restricted to individual EHR data formats [12]. This is recently changing as more and more solutions are appearing.

Canual et al.[13] analysed the following features for the seven precision medicine research platforms listed below: Information content (clinical and omics data), privacy management environment, analysis supports, visualization tools, interoperability support, system requirements, programming language and platform support [13]:

1. Biology-Related Information Storage Kit (BRISK) [14] which is an open source Web application providing access to phenotype and genotype data allowing researchers to conduct GWAS analysis;
2. Integrated Clinical Omics Database (iCOD) [15][16] that allows the researcher to collect and combines data pertaining to hepatocellular carcinoma (HCC) cases;
3. Integrating Data for analysis, Anonymization and SHaring (iDASH) [17] which is a computational collaborative cloud infrastructure conceived to share patient data for research;
4. tranSMART which is a software framework that allows the analysis of integrated data for the purpose of hypothesis creation, hypothesis validation, and cohort discovery needed in Precision Medicine;
5. Oncology Data Retrieval System (OncDRS) [18][19]. A system that query and integrates clinical and genomic data from heterogeneous sources;
6. Georgetown Database of Cancer (G-DOC) [13][19], A data integration and interrogation knowledge discovery system for oncology and precision medicine; and
7. cBio Cancer Genomics Portal for Cancer Genomics [19] [20], an open-access platform to explore cancer genomics data.

The first feature is whether the platform supports genetic data. We noticed that all platforms store the genetic information of a patient. The second feature assessed is whether the platform supports other data required for precision medicine research. We also conclude that all these platforms can store clinical and some can also store environmental data about a patient.

The third feature assessed concerns the data model extensibility: is the platform capable of supporting any other patient data requirement that does not exist in the current platform data model without incurring a major effort? We found that none of the platforms offered the possibility to easily adjust its proposed data model for specific precision medicine analysis needs. Fourth, we assessed a criterion about theIT infrastructure portability: is the precision medicine platform proposed easily portable to different cloud computing suppliers? Here again, we did not find any indication of the possibility to move across cloud suppliers. In some cases you are locked in the platform IT infrastructure, which does not explain where/who, operates it.

For the fifth feature,we tried to assess the data scalability

offered by the platform: does the platform allow the researcher to efficiently load very large amount of data without incurring any data model or infrastructure limitations?We concluded that the platforms were designed to allow for the input of a large volume of data about patients. The sixth feature assessed addresses the cloud infrastructure scalability: doesthe platform allow processing on a cloud distributed scalable infrastructure, and provide tools to scale as needed? We did not find any indication of the possibility to scale the cloud infrastructure as required. Finally, we were looking for a research reproducibility function: does the platform allow researchers to reproduce their research at any time without a great deal of effort.We found that most platforms offer this possibility.

II. PRECISION MEDICINE PLATFORMS FUTURE DIRECTION

The precision medicine platformof the future will need to offer all the features presented earlier and more. Its features should be available as SaaS (e.g.software as a service) and allow its operation on any cloud-computing supplier. It should also have been designed and programmed using open-sourceBig Data technologiesthat cheaply allowfast in-memory processingof large amounts of patient data. Hospital research labs will want to profit from the agility of a 'pay as you use'approach offered by cloud computing suppliers so that they do not have to wait and line up to run their analysis on university or government research supercomputers/clusters or buy theirown hardware and software. Most of all, a key ability required by the precision medicine researchers will be allow them to easily adjust the data schema of the database to their specific research needs as they evolve and change. Finally, they will want to scale the cloud IT infrastructure as needed to get their results as fast as they need them.

III. PROPOSING AN EXTENSIBLE PRECISION MEDICINE PLATFORM

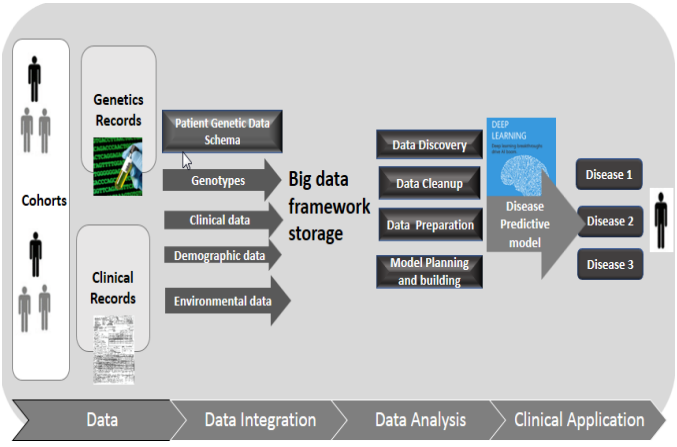

Fig. 1 Future Precision Medicine Platform

We have designed such a platform, with the help of AWS Canada researcher grant. This first prototype version of our precision medicine platform of the future has been designed in three main components:

1. **Input Data:** Allowsfor the combination of different data sources and their data storage. The main data sources are trial cohorts and electronic health care records. The volume of this data varies and increases continuously;
2. **Data integration:** Allows for the collocating of the required precision medicine research data in one single data model that can be implemented on a scalable cloud infrastructure;
3. **Data Analysis:** Researchers can use any data analysis tool to conduct their research analysis and mine this data.

In addition, the precision medicine platform of the future needs to automate the different data preparation steps (e.g. steps 1 to 8 of figure 2) involved in a typical Precision Medicine research activity:

1. Identification/collection of the data required for the research goals;
2. Mapping of the data fields to the existing data model to check if it contains all the data fields needed; otherwise, can be used the data creation model component to add the missing data elements;
3. Use the scalable data migration infrastructure setup component to configure the data migration computer infrastructure for the performance needed;
4. Start the data migration component to load the data from the many data sources into the integrated database;
5. If the volume of the data to be loaded is very large, such as for genetics, the researcher can easily scale up the computer infrastructure by adding new instances;
6. Once the data is loaded and ready for analysis, the researcher can use the data analysis infrastructure setup component to configure the infrastructure and environment required to start the analysis;
7. The researcher conducts his precision medicine analysis on the data;
8. Based on the analysis performance needed, the researcher uses the scale data analysis framework structure tool, at any time, to scale up the computer infrastructure to fit the performance requirements for their analysis goals.

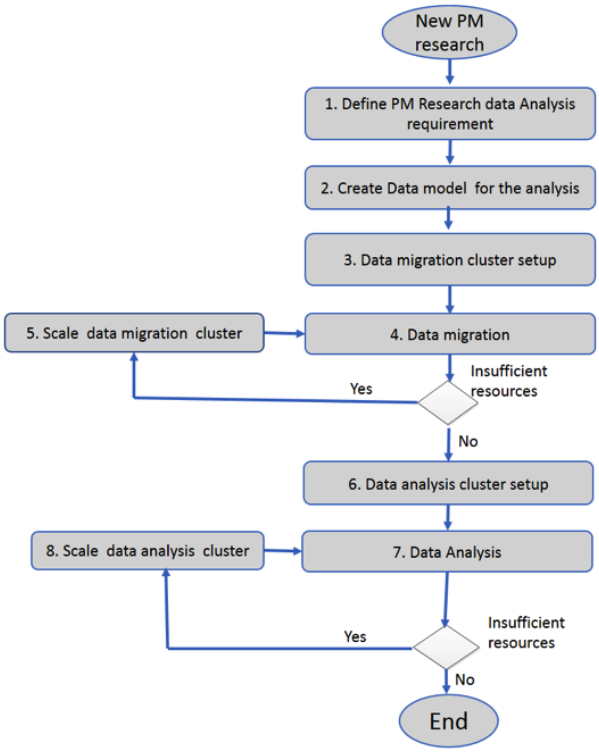

Fig. 2 Future Precision Medicine Data Analysis Pipeline

By automating these steps, researchers that do not have specialized IT skills available in their lab should be able to prepare their data by themselves in a few steps.

IV. PROPOSED FRAMEWORK SOFTWARE ARCHITECTURE

The proposed framework designed in our prototype includes a number of APIs, that use: 1) an already proven to be scalable data model; 2) open-source Big Data technologies, like the Apache Hadoop distributed file system (HDFS), Spark, Parquet and Yarn to ensure the scalability in processing high volumes of data; and finally 3) a scalable cloud computing infrastructure on the cloud (e.g. Azure, AWS or Google Cloud). Figure 3 shows how we have architectured each of these freely available open-source software components to meet our requirements.

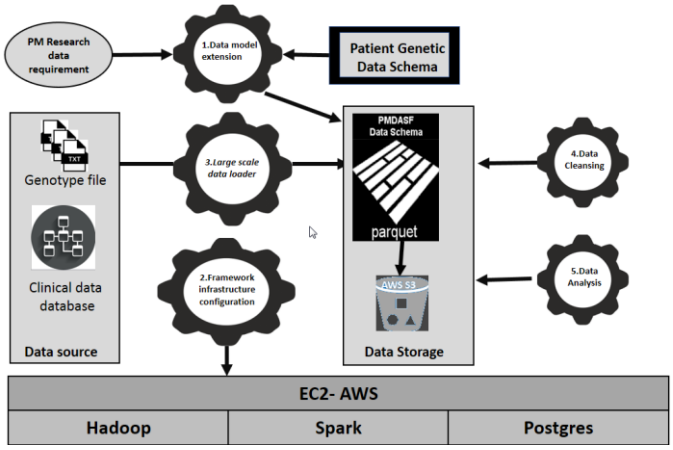

Fig 3. Proposed future precision medicine platform

V. PRELIMINARY RESULTS

To test this proposed platform design, we conducted an initial case study which involved preparing the data for a precision medicine study that concerned developing a predictive model for the complications of chronic kidney disease (CKD) in patients with type 2 diabetes using the patients' genetic variant, clinical and environmental data. It uses a list of informative genetic variants encompassing relevant risk factors for CKD complications, selected from publicly available GWAS data and tests them on the ADAVANCE cohort data [21], [23]. For this case study we have followed each step recommended by Figure 2.

The first step is to gather the data. With the researcher, we located the data of 1118 patients that were previously genotyped using Affymetrix's GeneChip arrays resulting in101 GB of data located in many individual files of .Gen format. Then we located and studied the clinical data of these patients that was stored in a Postgres relational database.

In step 2 of Figure 2, we proceeded with the data model adjustment for this specific precision medicine study. The default data schema for the genotyped data did not need any adjustment, as it is quite standard. Alternatively, we needed to adjust the database schema to add the clinical data as well as the analysis schema specific to this study (see Figure 4). It required to add the following five data classes: personal, visit, diagnostic history, phenotype and medical treatments. This data extension is required as the data items found in the Postgres databases comes from the ADVANCE trial case-cohort [21], [22] and needs to be added for our future analysis.

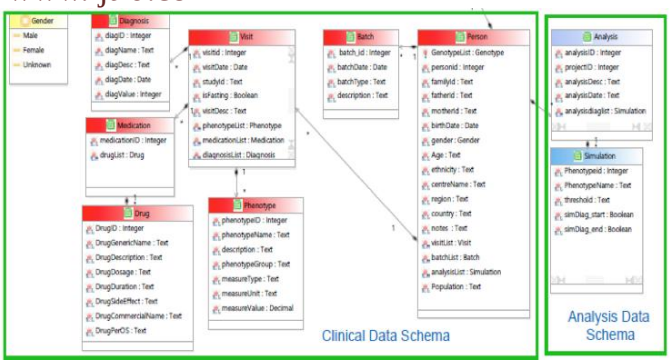

Fig 4. Added clinical and analysis data schemas

Since our goal is to enable a large-scale and complex precision medicine analysis of this patient data using Big Data and machine learning algorithms we need to collocate all this data in a centralized database schema. The colocation of all this data, in a single database, can allow precision medicine researchers to exploit the power of popular open source Big Data technology(e.g. Hadoop, Spark, Avro and Parquet) cheaply to try to identify potential correlations such as candidate genes responsible for specific diseases and impact of therapies and medications on a patients' health. This precision medicine analysis is typically evaluated using a patients' genetic data, clinical data as well as other data such as age, gender, ethnicity and weight. Another data schema extension is needed and presented in the box on the right side of Figure 4. Its aim is to allow researchers to better organize, track, and reproduce the results of an analysis. It is composed of the following two main data classes: Analysis and Simulation. The whole process, of the database schema extension, took only 2 minutes using our prototype data creation model component API that was designed to easily add the missing data elements.

In step 3, we setup the cloud infrastructure for the data conversion/loading process.It took less than 8 minutes to configure the Amazon Web Services (AWS) cluster of servers. Using AWS for our case study, we configured a cluster of 10 Linux instances (m4.4xlarge) ready for use. The configuration allocated 144 cores and 557 GB of memory in total.

In step 4, our prototype platform API's are used to convert/load the data into the target database schema. First, the genetic data of each of the 1118 genotyped patients (101 GB) needed to be converted from the Oxford genotype file format (.Gen) generated by the Affymetrix GeneChip arrays. This process took 3 hours 18 minutes. Second, the ADVANCE clinical needed to be extracted from its Postgres relational database (71 MB) and loaded in the target database. This process took only 1.4 minutes.

Next we conducted some extra experimentation to demonstrate the utility ofthe step 5 of Figure 2. We proceeded to scale up the processing of the patient genetic migration step. We noticed that it was possible, in only 50 seconds, to add an instance to the AWS cluster. For amore powerful infrastructure, it was also easily possible to add 5 instances but this took 3 additional minutes to be executed. This experiment concluded that a researcher that is unhappy with the 3.3 hours wait for loading large genetic files can quickly adjust the power of the cloud infrastructure as required to reduce this time. For this first case study, we did not go further in our experimentation.

Our last research taskof this case study was to calculate the cost associated with thisprecision medicine data preparation. We were pleasantly surprised to find that the total cost for: 1)

the data schema extension(2 min.); 2) the cloud cluster infrastructure setup (8 min.); 3)the data conversion/migration (genetic data 3 h.18 min. and clinical data 1.4 min.) was only 32$ USD at 0.80$USD/hour for each instance (in this test we used 10 m4.4xlarge instances). Next, add to this the cost the storage of 101.07 GB of data costing only $2.32/month.Last, there is another small fee to consider in this budget. The AWS requests price (e.g. Get and Put at $.005 per 1,000 requests) that did not go over $4.

Now, in less than 4 hours and for this very low cost under $40 USD, the database is ready for a large-scale precision medicine analysis.

CONCLUSIONS

Novel precision medicine platform designs should allow researchers to easily adjust the data model and scale the data preparation/loading on demand. It should also allow the analysis activities to be conducted on a single database designed for lightning fast and scalable processing. In this paper, we presented a proposed design that includes popular Big Datatechnologies and stepsto easily prepare the data for any research involving a patient's genetic, clinical and environmental data. A prototype of this platformwas experimented and demonstrated a number of advantages: the possibility to easily and quickly adapt the data schema for any precision medicine analysis requirement; a simple process to prepare the infrastructure for converting/loading large amount of genetic and clinical data; and the possibility to scale up the cloud supplier cluster infrastructure when needed. Finally we showed the low cost associated with the preparation of this large and complex data.

NEXT STEPOF THIS RESEARCH PROJECT

In our next publication, we describe how this framework was implemented and trialled by the precision medicine researchers, at the Centre Hospitalier Universitaire de l'Université de Montréal (CHUM), to conduct step 7 (data analysis) a large-scale precision medicine analysis of diabetes patients usinga clinical data set including 2394 patients anda their genetics data set (15,213,486,960 rows) as well as the Single nucleotide polymorphisms (SNP) list associated with eGFR gene and the urinary albumin to creatine ration (ACR)risk group. The data that was reshaped into one single data table comprising 112 columns: where 76 columns were used for genetic data and 36 columns represented clinical data (i.e. age, gender, region) across 1118 rows (i.e. 1 row per patient). The analysis included 10 simulations (iterations) with each of the following three classification models: logistic regression, random forest and neural networks to develop a CKD predictor.

References

[1] J. Davis-Turak*et al.*, (2017) "Genomics pipelines and data integration: challenges and opportunities in the research setting," *Expert Rev. Mol. Diagn.*, vol. 17, no. 3, pp. 225–237.

[2] J. L. Vassy, B. R. Korf, and R. C. Green, (2015) "How to know when physicians are ready for genomic medicine," *Sci Transl Med*, vol. 344, no. 6188, pp. 1173–1178.

[3] R. Gullapalli, M. Lyons-Weiler et al., (2012) "Clinical Integration of Next Generation Sequencing Technology," *NIH Public Access*, vol. 32, no. 4, pp. 585–599.

[4] K. Y. He, D. Ge, and M. M. He, (2017) "Big data analytics for genomic medicine," *Int. J. Mol. Sci.*, vol. 18, no. 2, pp. 1–18.

[5] J. Larry Jameson and D. L. Longo, (2015) "Precision Medicine—Personalized, Problematic, and Promising," Obstet. Gynecol. Surv., vol. 70, no. 10, pp. 612–614.

[6] D. J. Duffy, (2016) "Problems, challenges and promises: Perspectives on precision medicine," Brief. Bioinform., vol. 17, no. 3, pp. 494–504.

[7] A. E. Guttmacher and F. S. Collins, (2002) "Genomic Medicine — A Primer," N. Engl. J. Med., vol. 347, no. 19, pp. 1512–1520.

[8] A. M. Feldman, (2015) "Bench-to-Bedside; Clinical and Translational Research; Personalized Medicine; Precision Medicine-What's in a Name?," Clin. Transl. Sci., vol. 8, no. 3, pp. 171–173.

[9] E. M. Goldblatt and W.-H. Lee, (2010) "From bench to bedside: the growing use of translational research in cancer medicine.," Am. J. Transl. Res., vol. 2, no. 1, pp. 1–18.

[10] P. Garrido et al., (2017) "Proposal for the Creation of a National Strategy for Precision Medicine in Cancer: A position statement of SEOM, SEAP and SEFH," Farm. Hosp., vol. 41, no. 6, pp. 688–691.

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

[12] S. J. D. Chloé Cabot, Lina F. Soualmia, (2015) "Intégration de donnéescliniques 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.

[13] V. Canuel, B. Rance, P. Avillach, P. Degoulet, and A. Burgun, (2015) "Translational research platforms integrating clinical and omics data: A review of publicly available solutions," Brief. Bioinform., vol. 16, no. 2, pp. 280–290.

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

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

[16] L. A. D. Alessandro, R. Meyer, and U. Klingmüller, (2013) "Hepatocellular carcinoma : a systems biology perspective," vol. 4, no. February, pp. 1–6.

[17] L. Ohno-machado et al., (2012) "iDASH : integrating data for analysis , anonymization , and sharing," pp. 196–201.

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

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

[20] P. Tarczy-Hornoch et al., (2013) "A survey of informatics approaches to whole-exome and whole- genome clinical reporting in the electronic health record," Genet Med, vol. 15, no. November 2012, pp. 824–832.

[21] S. R. Heller, (2009) "A summary of the ADVANCE Trial.," Diabetes Care, vol. 32 Suppl 2, pp. 1–5.

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

[23] ADVANCE Management Committee, (2001) "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.