Clinical and genomics data integration using meta-dimensional approach

Conference Paper · December 2016

DOI: 10.1145/2996890.3007896

Moeez Subhani

University of Derby

1 PUBLICATION **1** CITATION

SEE PROFILE

Ashiq Anjum

University of Derby

98 PUBLICATIONS **2,238** CITATIONS

SEE PROFILE

2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing

Clinical and Genomics Data Integration using Meta-Dimensional Approach

Moeez M. Subhani
Department of Computing and Mathematics
University of Derby
Derby, England
M.subhani@derby.ac.uk

Ashiq Anjum
Department of Computing and Mathematics
University of Derby
Derby, England
A.anjum@derby.ac.uk

Andreas Koop
Diagnostics Global Informatics
F. Hoffmann-La Roche AG
Basel, Switzerland

Nick Antonopoulos
Department of Computing and Mathematics
University of Derby
Derby, England

ABSTRACT

Clinical and genomics datasets contain humongous amount of information which are used in their respective environments independently to produce new science or better explain existing approaches. The interaction of data between these two domains is very limited and, hence, the information is disseminated. These disparate datasets need to be integrated to consolidate scattered pieces of information into a unified knowledge base to support new research challenges. However, there is no platform available that allows integration of clinical and genomics datasets into a consistent and coherent data source and produce analytics from it. We propose a data integration model here which will be capable of integrating clinical and genomics datasets using meta-dimensional approaches and machine learning methods. Bayesian Networks, which are based on meta-dimensional approach, will be used to design a probabilistic data model, and Neural Networks, which are based on machine learning, will be used for classification and pattern recognition from integrated data. This integration will help to coalesce the genetic background of clinical traits which will be immensely beneficial to derive new research insights for drug designing or precision medicine.

Keywords

Clinical data; Genomics data; Data Integration; Meta-dimensional; Bayesian Networks; Neural Networks.

1. INTRODUCTION

Research in medical science is ample and widespread. Traditionally, the research was mainly focused on comprehension and contemplation of the clinical data points collected from medical practices. Clinical data points from different datasets were studied in different combinations, such as combining blood test results with physical examination reports and patient history to understand the nature and effects of a disease for a particular blood group. The effects on a disease due to genetic make-up of a person could not be captured due to unavailability of any genomics datasets. Since the advent of genome sequencing technology, the spectrum of data points available for research has developed extensively and beyond clinical datasets only. With the rapid development in genomics tools and technologies, there are huge amounts of genomics data now available for research and analysis.

The genomics datasets provide information about the genetic make-up of a person. The clinical features of a person are eventually the physical expression of their underlying genetic structure. The genotype is the genetic make-up of an individual, and the phenotype is the physical expression of a gene or interaction of different genes in an organism [1]. Essentially, associating a phenotype with its genotype can reveal the underlying genomic polymorphism for a disease or treatment.

To better understand the physiological features from clinical traits, it is crucial to relate them to their genetic architecture. In recent years, an increasing interest has been observed in research to combine clinical and genomics datasets [1][2][3]. Most of the research has been focused on finding associations between clinical and genomics parameters from single datasets only, such as associating gene expressions data and single nucleotide polymorphism (SNP) data [4][5]. The association studies are particularly common in precision medicine, also referred to as personalised or genomic medicine in research environment.

The challenge of associating or combining different datasets together makes it a data integration problem. The data integration process can be defined differently in different data science frameworks. In simple words, it is a problem of combining data from different sources and providing a single unified access to the data [5][7]. Data integration can also be defined as a means to query across different data sources [1][8]. Although data integration and precision medicine are separate domains of research; the two need to evolve together since the pressing challenges in medical science can only be solved through data driven discoveries [1][7].

In clinical environment, the data sources can be greatly diversified, such as health records, clinical trials and disease records etc. On the other hand, in genomics environment, the sources are very data intensive such as genome sequences, variants, annotations and gene expressions etc. Therefore, a new data integration model is needed that can take into consideration variety, volume and semantics of these data sources into a unified data source that can then lay foundation for new medical discoveries.

The major benefit of integrating clinical and genomics data will be to support precision medicine. Combining genomics and clinical parameters will provide genetic background of clinical problems at a more explicit level. Having genomic information of patient integrated with their clinical data will help medical practitioners to design more personalized treatment plans. Also, the pharmacogenomics industry will be able to provide more

UCC '16, December 06-09, 2016, Shanghai, China.
© 2016 ACM. ISBN 978-1-4503-4616-0/16/12…$15.00.
DOI: http://dx.doi.org/10.1145/2996890.3007896

personalised solutions, such as designing drugs with improved efficacy. Researchers will also be able use the integrated data to discover the insights behind complicated biological problems, such as designing new drugs or finding new biomarkers. Overall, the integration of data will benefit everyone related to these dimensions of medical science.

In this paper we propose a data integration model to integrate the genetic backbone with clinical data skeleton. We propose a meta-dimensional integration approach to integrate clinical and genomics datasets. This paper will highlight the initial phase of this PhD project mentioning the challenges about integration, the state of the art and the proposed solution.

2. RESEARCH PROBLEM AND CHALLENGES

The first and foremost challenge of this research work is to address the problem of data integration. The datasets under consideration in this study are structurally and characteristically very different due to their different origins.

The clinical datasets are comprised of various data types, such as health records, diagnostics test results, disease histories, digital scans and clinical trials datasets [10]. Most of these datasets are in tabular formats containing integers or text values, while others can contain images, such as X-Ray scans [11][12]. The number of parameters within each of these datasets is large and diverse. The variation in parameters increases further across different studies. For example clinical trials conducted by different organisations may capture different types of variables due to the underlined requirement of their respective studies, or health records stored by different hospitals or clinics may store information in different formats. Hence, the diversity within clinical datasets, in terms of parameters or data types, is generally large. These datasets are collected and stored mainly by hospitals, non-profit organisations or pharma industry in their respective data repositories or warehouses, commonly referred to as clinical data repositories (CDRs) [10][11]. The access of these datasets is not always directly available due to the personalised information of patients involved. Hence, getting access to all types of clinical datasets is another problem.

The genomics datasets are available in completely different formats than the clinical data. The genomics data is generated from genome sequencing technologies. There are several formats in which gene sequence data is stored, such as FASTA, FASTQ, GCG, XML, ASN.1 and GenBank etc. [13]. The information about variations among genes is stored in variant files, known as VCF (variant call format) files [14]. This data is stored in various repositories, such as NCBI, EBI, Ensembl, 1000genomneproject, GenBank, GeneDB are a few to name. A long list of other genomics databases is available worldwide, whereas, the protein databases are supplementary to them.

The prime challenge, therefore, lies in the integrability of these different data formats. The problem being addressed in this paper is about data integration of different data types and formats, which are intrinsically diverse and massive. The interoperability of data formats needs to be addressed by either transforming them into a common compatible format, or creating a platform that can accommodate various data formats.

The overall research challenge is to empower clinical research with genomics information. The integrated data should be capable of uncovering unidentified relationships among datasets, thus providing greater insights across the data space. This research project will explore the possible research insights within the integrated data which are not explicable independently, i.e. the information which is interdependent and can be revealed by data integration only. For example, combination of which genes are mutually affecting a particular disease type in a specific cohort of patients, or discovering a drug which is appropriate for a patient with a peculiar set of physiological features and a definitive gene set.

In particular, this research work will address the following research questions:

1. How to integrate clinical and genomics datasets to make them more useful and propitious for medical research?
2. What clinical and genomics data parameters needs to be integrated in order to connect important pieces of information?
3. How to analyse the integrated data in order to solve a biological problem, such as finding or validating a biomarker?

3. STATE OF THE ART

The existing data integration solutions combine clinical datasets only. These solutions are targeted for the purpose of combining various clinical datasets from different sources and making them available from a single platform. Currently, clinical data integration solutions are being provided by vendors like SAS [15], Lumeris [16], Edifecs [17] and others. These solutions are only meant to combine disparate clinical sources. None of the current solutions provide so far the mechanism for clinical and genomics data integration.

Due to absence of any data model that can accommodate both clinical and genomics datasets, there is a need to construct a data model which is capable of providing a single platform to integrate both kinds of datasets. Since the last decade research has increasingly steered towards this direction [1][3]. Researchers have explored various integration models for the sake of integrating multi-omics data, both clinical and genomics. The two most common approaches that can be found in literature are multi-stage analysis and meta-dimensional analysis [1][2][3]. The taxonomy of the two approaches has been compiled together as shown in Fig. 1.

Multi-stage analysis is a stepwise or hierarchical analysis method. It helps to reduce search space by using step-by-step analysis [2]. It essentially integrates and analyses only two data types at a time, while functioning across the data space. Triangle method is the most common method under this approach which has been widely used for association studies. This method is more commonly used for SNP associations along with gene expression data and genes themselves [4][18].

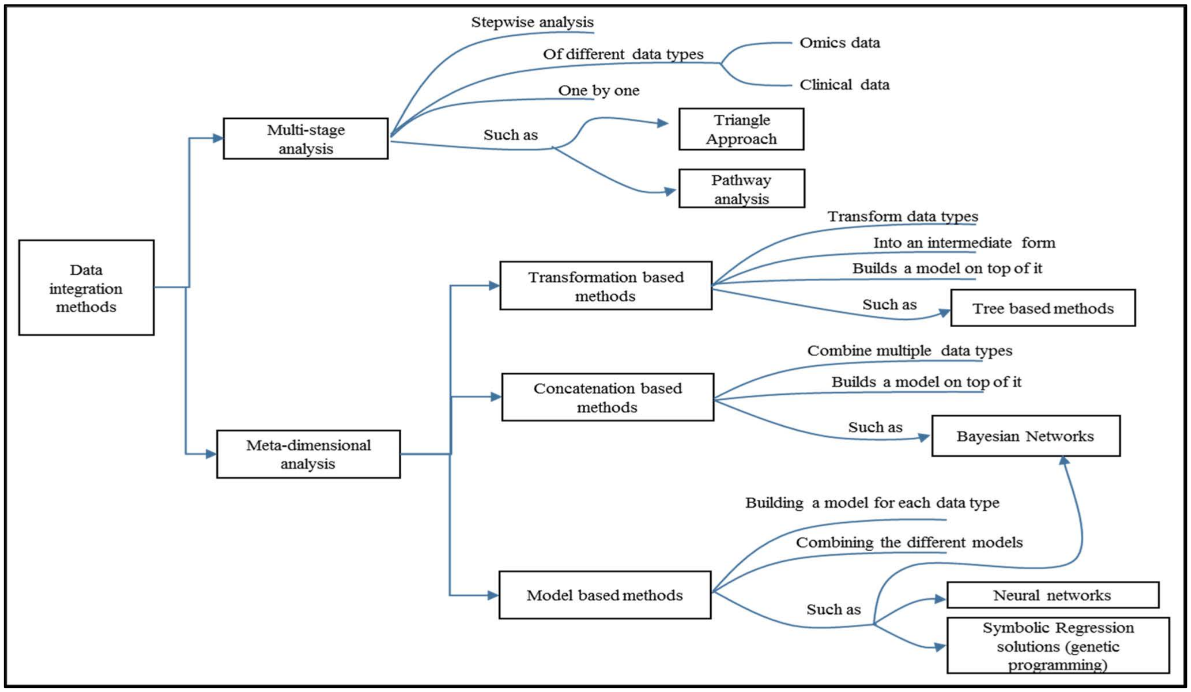
Data integration methods
Multi-stage analysis
Stepwise analysis
Of different data types
Omics data
Clinical data
One by one
Such as
Triangle Approach
Pathway analysis
Meta-dimensional analysis
Transformation based methods
Transform data types
Into an intermediate form
Builds a model on top of it
Such as
Tree based methods
Concatenation based methods
Combine multiple data types
Builds a model on top of it
Such as
Bayesian Networks
Model based methods
Building a model for each data type
Combining the different models
Such as
Neural networks
Symbolic Regression solutions (genetic programming)

Figure 1 Taxonomy of Data Integration Methods

Some physiological features are a result of interaction of different genes and physiological parameters. Due to step-wise analysis, this approach is not capable of capturing such features which are influenced by factors coming from different genes. Although it is robust and a rather simple approach, it is not recommended when multiple sources are required to be integrated [2][3].

Meta-dimensional approach, on the other hand, involves simultaneous analysis of multiple data sources to produce complex models [2]. This approach relies on building a new data model on top of data sets. There are various methods under this approach, such as concatenation based, transformation based and model based methods. The variation lies at what stage the new model is being built on the data. Fig. 1 shows the different methods within this approach. The meta-dimensional approach facilitates the provision of search across entire data space, allowing detection of those clinical features as well as features which are caused by mutual interaction of multiple physiological and genetic factors. However, the integration leads to complex and less robust models [2][3]. Among various methods within this approach, as illustrated in Fig. 1, Bayesian Networks and Neural Networks appear to be common in the integration based research [19][20].

Some analysis platforms for clinical and genomics data are also available, such as ATHENA and tranSMART. ATHENA (Analysis Tool for Heritable and Environmental Network Association) is a multifunctional analysis tool which provides a platform to apply statistical techniques and identify meta-dimensional models [21]. It provides filtering and modelling components, and evolutionary computing approaches to generate prediction models from datasets. The results from ATHENA are promising; however, it is limited only to the modelling parameters it provides. Currently it supports only two evolution models: GENN and GESR [22].

tranSMART is a web-based knowledge management platform [23]. It essentially provides scientists and researchers a platform to validate their results and hypotheses by investigating the correlations between phenotypes and genotypes [24]. tranSMART facilitates association studies, provides search, compare and contrast functions on datasets and displays data visually by using graphical interface. Essentially, tranSMART can be used as a platform to verify the results, but it cannot be integrated with external (non-web based) data models.

To summarise, there is a strong need to provide a platform which can integrate clinical and genomics datasets and facilitate analysis on them to discover insights. Meta-dimensional approach appears to be more promising for this task since it can serve for multiple datasets simultaneously, which is essential in this case. One of the meta-dimensional approach based methods can be adapted, modified and extended to build a data model according to the structure of datasets and needs of this project. Neural Networks provide very efficient, robust and intelligent method to derive insights from data, and therefore, they can be used to derive covert insights from integrated data. A validation source will be required afterwards to verify the results. tranSMART can be a good source but it doesn't support integrated datasets currently. An alternative source or a hybrid source needs to be determined for validation purpose.

4. APPROACH AND SCOPE OF STUDY

The aim of this study is to explore the insights of clinical datasets by integrating them with genomics datasets. The overall approach will be to extend existing data models that store clinical datasets by including clinical trials, health records as well as genomics datasets, under the same roof. A conceptual extended data model is illustrated in Fig. 2.

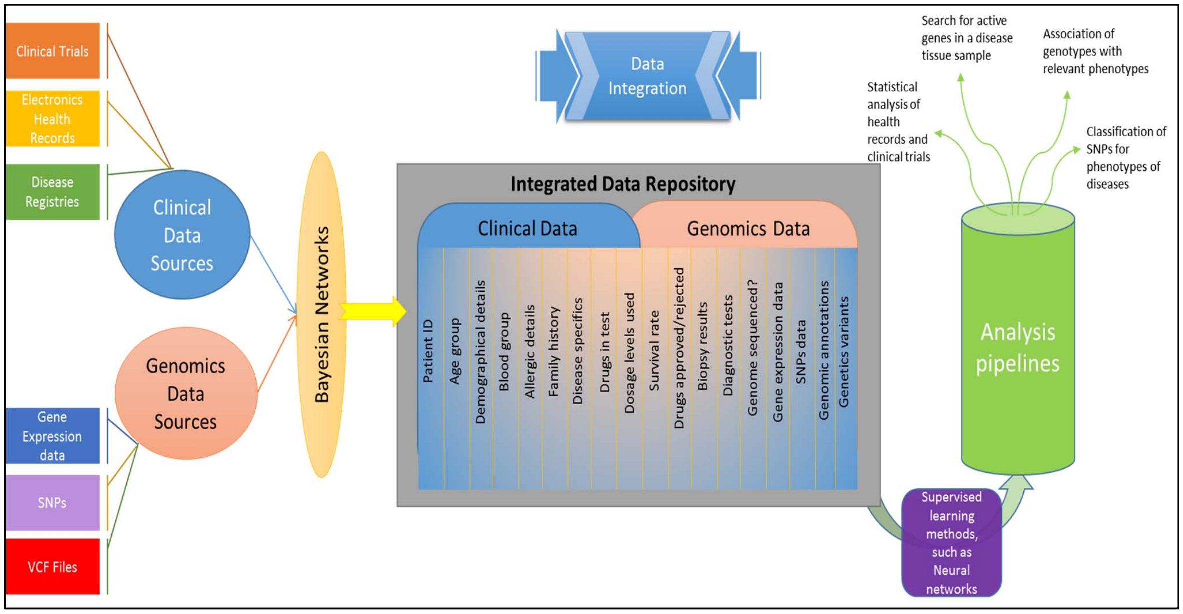

Figure 2 Conceptual Data Model

The datasets being integrated should be relatable and compatible with each other in terms of their structure and the information they contain. A meta-dimensional approach will be attempted to build an integrated data model and implement it in a data warehouse to address the issues described in the problem statement. This implemented data model should be scalable to accommodate big datasets while keeping the same performance.

The scope of this study will span data integration and data analysis. The data integration model will be designed in a more generalised context in order to accommodate any type of data within the clinical and genomics data domains. In addition, the integration has to be dealt intelligently so that the resulting data set can enumerate value to the clinical research. Similarly, the analysis of integrated data has to be performed such that it will find data elements from individual datasets, and endeavour the covert insights and information for wider medical research.

The data integration model will be novel in itself because no such data integration model exists which can encompass such a variety of datasets. As mentioned previously, the state of the art solutions do not provide this functionality where clinical and genomics datasets can be brought together at a single platform and perform analysis on them. The effort has been made only up to integrating single domain datasets [5][9][18][19]. However, there is a wide gap in this research area which needs to be fulfilled to achieve clinical and genomics data integration.

5. PROPOSED SOLUTION

Based on findings from literature review and considering the needs of this project, the solution proposed for this problem will be based on meta-dimensional approach. The solution will exploit the advantages of Bayesian Networks and Neural Networks methods in order to integrate and analyse, respectively, clinical and genomics data. Since Bayesian Networks can be used under both concatenation and model based methods, it provides more flexibility in terms of implementation. A similar approach has been used previously for an association analysis by Fridley et al. to design an integrative genomic model [19].

Bayesian Networks (BN) are probabilistic graph models, which represent a set of random variables and their conditional interdependencies. They are essentially directed acyclic graphs, and have been seen to be popular in the fields of statistics and artificial intelligence [25]. A typical BN structure is characterised by a set of nodes and edges. The nodes represent random variables, and edges represent the relationship between nodes. Since the edges are directed, they represent the probabilistic dependencies among the variables [25]. The arrows on the edges essentially represent the 'influence' of a node over others, which statistically means that the value taken by a node is a result of the influence (or is calculated) from all the other nodes directed towards it. Henceforth, if node A is directed toward node B, then node A is referred to as parent node for node B, or conversely node B is a child of node A [25]. In our case, each node will represent a parameter, from either clinical or genomics datasets, and the edges connected to this node will represent the relationship of this parameter to other connected nodes (parameters).

The datasets under consideration here comprise of a large number of parameters with various data types. These parameters may have different dependencies over each other. To infer the correct semantics of the parameters, the probability distribution of each parameter needs to be determined. A probabilistic graph model approach allows to perform these probabilistic queries across datasets and infer their semantics [26]. The famous classic Markov Models, and its variants, also allow to calculate the conditional probability distribution for each node, which physically establishes the probabilistic relationship of a node to its parent nodes [27]. However, in case of clinical and genomics datasets, the interdependencies among variables will be very complicated, and in addition to determining the conditional probability distribution, joint probability distribution needs to be asserted as well to establish the relationship of a node in the given probability space. The probabilistic graph models allow to calculate both conditional and joint probability distributions, and, therefore, seem more appropriate for this study.

These graph based probabilistic models can be compiled with relational schemata to construct a probabilistic relational model for relational datasets. BN being probabilistic graph models can be used to create a probabilistic relational model [26][27]. Star schema is one of the relational schemas common in literature. It has been adopted and modified previously by Zhang and group for integration of clinical and genomics datasets [28][29]. A typical star schema consists of a single central fact table and multiple de-normalised dimension tables [21]. Due to large number of variables within clinical data space, the numbers of parameters and their interdependencies are also high. Many-to-many relationships between fact and dimensional tables are common in both clinical and genomics datasets. For example, a single gene can influence multiple phenotypes and similarly a single phenotype can be a result of interaction of multiple genes. The original star schema is mainly designed to handle many to one relationships; hence, this model needs to be extended to meet the requirements of our data space. Biostar schema, designed by Wang et al., has addressed some of the issues to cover both clinical and genomics data types [28].

The schema to be designed for this work will be a probabilistic relational schema. The biostar schema will be modified and extended to probabilistic schema so that it can be mapped over Bayesian Networks based data model. Since biostar schema is particularly tailored to target biomedical datasets, it will be a fitting option to capture all types of clinical and genomics datasets. Extending a relational schema to a probabilistic schema means to modify it to include the probability distributions calculated from the BN model [26].

The data model and schema will be implemented in an SQL database. Since the incoming data is already structured, it will be reasonable to push it into an SQL database warehouse, and we limit scope of this study to SQL databases only. A conceptual data model is proposed in Fig. 2.

The integrated data will be used for further analysis, such as classification, prediction or pattern recognition. Neural Networks offer all these analytical functionalities and have capability to find associations among complex variables and discover hidden complicated relationships among them. Therefore, they can be very effective and useful to perform machine learning on the integrated data. Although, an exact analysis scheme will be designed at a later stage, a rough idea is to employ various types of Neural Networks for various analytical needs, such as multi-layer or convolutional Neural Networks. The main idea is to classify the integrated data based on unique phenotypes and then further explore the hidden patterns for each phenotype. A simple basic 2-layer Neural Network can be designed for classification purpose. Similarly, there are various approaches based on Neural Networks and genetic algorithm for pattern recognition or forecasting.

To validate the solution, we will integrate clinical and genomics datasets to find predictive biomarkers for colorectal cancer. Relevant clinical and genomics datasets for colorectal cancer will be acquired primarily from public sources and imported to HANA based platform. The probabilistic data model and schema will be implemented to integrate datasets. Unique phenotypes will be identified from the integrated sets, using Neural Networks based pattern recognition methods, and then data will be classified based on them. Furthermore, association test will be performed to associate the phenotypes to their genotypes and discover the obscured relationships and establish predictive biomarkers from them. tranSMART [23] or Open Targets [30] can be used to verify the association test results. This case study will validate the whole proposed solution, and also demonstrate the scalability and performance of the integrated platform.

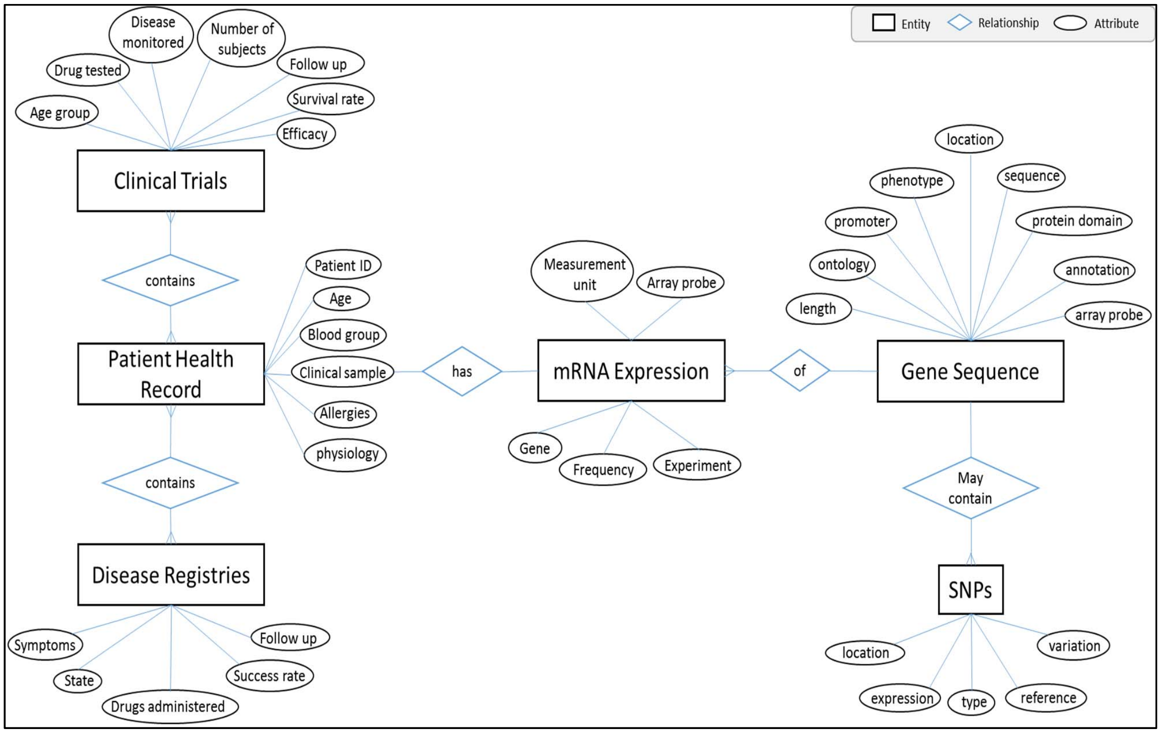

Figure 3 Entity-Relationship Diagram of Proposed Data Model

6. PROGRESS ACHIEVED AND FUTURE DIRECTIONS

In the first stage, the task of determining which clinical and genomics parameters need to be integrated has been achieved. Due to huge size of genomics datasets, and large variability in clinical parameters, it is not viable to integrate all parameters. Therefore, based on the literature research, gene expression data and SNP datasets have been identified to integrate with the clinical data initially. Determining the gene expression of SNPs can help to find out ultimate effects of a gene on a phenotype, therefore, these parameters have been selected initially to be integrated with clinical data. Further parameters may be added for integration at a later stage depending on the needs of an analysis scenario.

In parallel an entity relationship model, as shown in Fig. 3, has also been designed for the data model, which is essentially a first step toward designing a data model. This model illustrates how the link between clinical and genomic parameters can be established. The clinical trials data repositories contain health records obtained from hospitals, which contain the patient specific data. The genomics details of any patient is linked to patient records via gene expression data, given that the genome of the patient has been sequenced. Further information about the genome, such as SNPs data, can be obtained by performing annotations on sequence data using public sources. The BN based data integration mode and an appropriate schema will be built based on this entity relationship model. The project will be implemented in SAP HANA environment, and an instance has already been set up.

The research work is in progress and after having proposed a conceptual data model (Fig.2), and an entity-relationship diagram for it, the next task is to design and implement a schema for this model within SAP HANA and import the datasets. Public data sources will be used at initial stage for the implementation and analysis of the proposed data model. The next task will be to propose a meta-dimensional data model and a schema for it.

7. REFERENCES

[1] Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. and Tarczy-Hornoch, P., 2007. Data integration and genomic medicine. *Journal of biomedical informatics*, *40*(1), pp.5-16.

[2] Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. and Kim, D., 2015. Methods of integrating data to uncover genotype-phenotype interactions.*Nature Reviews Genetics*, *16*(2), pp.85-97.

[3] Hamid, J.S., Hu, P., Roslin, N.M., Ling, V., Greenwood, C.M. and Beyene, J., 2009. Data integration in genetics and genomics: methods and challenges. *Human genomics and proteomics*, *1*(1).

[4] Nevins, J.R., Huang, E.S., Dressman, H., Pittman, J., Huang, A.T. and West, M., 2003. Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human molecular genetics*, *12*(suppl 2), pp.R153-R157.

[5] Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C. and Lum, P.Y., 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, *37*(7), pp.710-717.

[6] Lenzerini, M., 2002, June. Data integration: A theoretical perspective. InProceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246). ACM.

[7] Doan, A., Halevy, A. and Ives, Z., 2012. Principles of data integration. Elsevier.

[8] https://www.techopedia.com/definition/28290/data-integration

[9] Orechia, J., Pathak, A., Shi, Y., Nawani, A., Belozerov, A., Fontes, C., Lakhiani, C., Jawale, C., Patel, C., Quinn, D. and Botvinnik, D., 2015. OncDRS: An integrative clinical and genomic data platform for enabling translational research and precision medicine. *Applied & translational genomics*, *6*, pp.18-25.

[10] https://en.wikipedia.org/wiki/Clinical_data_repository

[11] https://wiki.nci.nih.gov/display/TCGA/Clinical+Data+Overview

[12] Gilchrist, J., Frize, M., Ennett, C.M. and Bariciak, E., 2011. Performance evaluation of various storage formats for clinical data repositories. IEEE Transactions on Instrumentation and Measurement, 60(10), pp.3244-3252.

[13] https://www.genomatix.de/online_help/help/sequence_formats.html

[14] https://faculty.washington.edu/browning/beagle/intro-to-vcf.html

[15] https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-clinical-data-integration-103961.pdf

[16] http://lumeris.com/wp-content/uploads/2014/05/Lumeris-SOL.CDI_.05-14.v1.pdf

[17] https://www.edifecs.com/downloads/Clinical_Data_Integration_Solution_Brief_2015.pdf

[18] Lee, E., Cho, S., Kim, K. and Park, T., 2009. An integrated approach to infer causal associations among gene expression, genotype variation, and disease. *Genomics*, *94*(4), pp.269-277.

[19] Fridley, B.L., Lund, S., Jenkins, G.D. and Wang, L., 2012. A Bayesian integrative genomic model for pathway analysis of complex traits. *Genetic epidemiology*, *36*(4), pp.352-359.

[20] Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A. and Pe'er, D., 2010. An integrated approach to uncover drivers of cancer. *Cell*, *143*(6), pp.1005-1017.

[21] Holzinger, E.R., Dudek, S.M., Frase, A.T., Pendergrass, S.A. and Ritchie, M.D., 2013. ATHENA: the analysis tool for heritable and environmental network associations. Bioinformatics, p.btt572.

[22] Kim, D., Li, R., Dudek, S.M. and Ritchie, M.D., 2013. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData mining, 6(1), p.1.

[23] http://transmartfoundation.org

[24] Athey, B.D., Braxenthaler, M., Haas, M. and Guo, Y., 2013. tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. AMIA Summits on Translational Science Proceedings, 2013, p.6.

[25] Ben-Gal, I., 2007. Bayesian networks. Encyclopedia of statistics in quality and reliability.

[26] Singh, S. and Graepel, T., 2012. Compiling relational database schemata into probabilistic graphical models. arXiv preprint arXiv:1212.0967.

[27] Getoor, L., 2006. An Introduction to Probabilistic Graphical Models for Relational Data. IEEE Data Eng. Bull., 29(1), pp.32-39.

[28] Wang, L., Zhang, A. and Ramanathan, M., 2005. BioStar models of clinical and genomic data for biomedical data warehouse design. *International journal of bioinformatics research and applications*, *1*(1), pp.63-80.

[29] Du, N., Guo, S., Mahajan, S.D., Schwartz, S.A., Nair, B.B., Hsiao, C.B. and Zhang, A., 2012. BioStar+: a data warehouse schema for integrating clinical and genomic data from HIV patients. *ACM SIGBioinformatics Record*, *2*(3), pp.6-16.

[30] https://www.opentargets.org