Abstract

Summary: I2EHR is an approach to analysing intergrated clinical and genomic data and identifying patterns that are not commonly identifiable in unstructured clinical data. We present I2EHR.

The analysis of clinical data is hindered by the unstructured format of the data as it was originally entered as a means of storage. This lack of uniformity has resulted in subsequent

Motivation: gain insight into the benefits of integrating with clinical data and relating the patterns of gene expression that may associated with certain clinical characteristics.

Results: This paper describes the methods to model the association of clinical observations with differential gene expression patterns using basic clinical records that is commonly available in electronic health records. This model was validated using synthetic patient data and integrating with gene expression profiles to develop a realistic and longitudinal model of disease pathology . Lastly, the model incorporates the genotype-phenotype associations to perform predictive modelling for disease outcome. These results suggest that EHR stored clinical data can be integrated with genomic data to perform predictive modeling and improve treatment and outcome.

Introduction

The widespread implementation of the electronic health record (EHR) has improved healthcare through reduced processing times and cost, improved quality of patient care and detailed clinical information through longitudinal storage \citealp{PheWAS}. Longitudinal clinical data can be useful for applied clinical research like identifying drug adverse effects or respositioning of drugs. The extension of genomic data in EHRs in approaches such as biobanks have demonstrated the potential for pattern identification possible with integration. Despite the substantial discoveries possible from integration, many challenges need to be addressed. For example, a lack of standardisation has limited the degree of cohort analysis that can be performed. For example, most current EHR interface do not provide a timeline or visual representation of patient wellness recordings. This results in cumbersome and time consuming review of the patient information and limits potential for pattern identification.

The development of software to overcome this is limited by the sensitivity and privacy concerns regarding patient data. Approaches such as the HL7 Fast Healthcare Interoperability Resources (FHIR) provide a framework and data standards to improve interoperability and data exchange in EHRs. To address the potential for analysis, approaches like ShinyFHIR have integrated FHIR data into R for patient and cohort analytics. This approach demonstrates the adaptability of their app to work on numerous FHIR servers and bridge the gap between the lack

of programming skills that has limited application development. Approaches to address the sensitivity of the data include Synthea, an approach the produce realistic patient data that can be used for health informatics when access to real data is not mandatory.

To address current challenges associated with the integration of clinical and genomic data, I designed a model to integrate clinical and genomic data and provide visual and timeline representations of patient recordings. The aims of this dashboard included a system that would (a) develop a dashboard to integrate clinical and genomic data (b) create simple predictive analytics using the patient and clinical data and (c) download data from GEO to integrate and provide proof-of-concept for combined clinical and molecular analytics.

Given the increasing ease and widespread use of genomic analysis, this approach could be easily applied to many hospital EHRs to provide useful statistical results for genomic research. The tool requires little expertise to use and by development on R, it can be deployed to all devices with little to no configuration. By providing an integrated system containing combined analytics for clinical and genomic data, we aim to improve the accessibility of combined analytics and a patient and cohort level and further improve the utilisation of longitudinal patient data.

Materials and Methods

Clinical data: The clinical data modelling diabetes patients was generated using modules from Synthea, a synthetic patient generator. The population consisted of 1000 individuals originating from Massachusetts, MA. The selection criteria specified that patients suffered from a metabolic disorder. The generated data contained CSV files for information on allergies, careplans, conditions, encounters, imaging studies, observations, organizations, patients and procedures. The patients files contains demographic information such as first name, last name, address and gender. Each value is artificial and prevents any privacy issues that arise from identification of anonymised health records.

The observations file contained the information useful for predictive and visual data representation such as height and weight. Standard wellness measurements were generated using the wellness module. Diabetes-specific wellness measurements such as insulin resistance and glycated haemoglobin were generated using 2 metabolic modules namely metabolic syndrome disease and metabolic syndrome care. Insulin resistance and Hb1Ac levels were selected as the defining features that would be used for subgroups in the analysis.

Insulin sensitivity is measured using the rate of glucose infusion during the last 30 minutes of a peripheral vein infusion of insulin. If insulin infusion required exceeds 7.5 mg/min or higher, the patient is insulin sensitive. Levels below 4.0 mg/min indicate resistance to insulin action. Levels of 4.0 to 7.5 mg/min are an early indicator of insulin resistance.

The HbA1c level refers to the glycated haemoglobin levels. Glycated haemoglobin (HbA1c) occurs when haemoglobin, the iron containing protein that allows efficient transport of 4 oxygen molecules. This joins with glucose to become 'glycated' and represents to average blood glucose over an 8-12 week period. Normal HbA1c levels are below 6% (< 42 mmol/mol), prediabetes levels are between 6-6.4% (42-47 mmol/mol) and diabetes levels are above 6.5% (48 mmol/mol =<) .

Genomic data: Genotypic data was extracted from the Gene Expression Omnibus. Montesanto et. al indicate that CAT, FTO and UCP1 are all associated with type 2 diabetes. Search each of these genes with type 2 diabetes directs me to the same series: GSE25462. To model the effect of medication on gene expression, the commonly prescribed drug Simvastatin 20 MG Oral Tablet (n = 46) was used in the search. The search term

```
\begin{verbatim}
type[All Fields] AND 2[All Fields]
AND diabetes[All Fields]
AND Simvastatin[All Fields])
AND ("Homo sapiens"[Organism]
AND ("50"[n_samples]:"100"[n_samples])
\end{verbatim}
```

was used to find gene expression variation associated with the disease and drug and the . The top result measured APOC3 levels in subgroups Type 2 diabetic and insulin-resistant but normoglycemic cohorts in skeletal muscle cells. The primary outcome of the search was to identify a suitable model for diabetes associated gene expression variation. Studies have suggested a linked between mitochondrial oxidative homeostasis and predisposition to diabetic vascular complications (Alberto Montesanto, 2018). Plasma APOC3 associates with almost half of HDL (Zvintzou E, 2017); HDL-associated apoA-I and lysosphingolipids protect against ATP reduction resulting from reverse cholesterol transport in mitochondria CR White, 2017). When sorting this by subset effect, TPM1 comes up as highest.

Genotyping was performed using [\ HG-U133 Plus 2]\ Affymetrix Human Genome U133 Plus 2.0 Array which, according to their website, provides the greatest accuracy and reproducibility of any microarray platform.

Development of integrated genomic and clinical data: The generated data contained CSV files contain information on allergies, careplans, conditions, encounters, imaging studies, observations, organizations, patients and procedures. The observations were merged with patients to put the name to each wellness and disease associated measurement. The medical coding or the associated observation description could then be used to compare

Applications

I2EHR contains 5 major components that were deemed most relevant for the integrated analysis of the data: *introduction and project plans, patient clinical data, cohort genomic data, cohort clinical data, cohort genomic data.* The components of the clinical data components includes generating a barchart which displays the patients by ethnicity frequency, a hemoglobin measurements chart, the frequency of each disease subgroup, and the patient's BMI measurements over time.
The patient genomic data includes the gene expression profile associated with the individual. The features available for the analysis of the clinical data on a cohort level, include a report of the genomic data, the relative log expression, a principal component analysis, intensity filtering and heatmap samples, multidimensional scaling and microarray expression density. It also includes a plot of the H1Ac levels and a characterisation of the levels by disease prognosis category. The level of insulin resistance was also used as an indicator as to whether they were diabetic, prediabetic and normal.

1. ***Landing Page***
   The initial landing page indicated the project plan in a flowchart and gives information on individual data sources and generation. Contact information and a dashboard index is also available here.

2. Patient Clinical Data

On this part of the app, the user is prompted to enter the patient's ID. The ID number is used instead of the name to mimic the data protection actions that are generally put into place. Entry of the patient's name will then provide the user with the patient information such as their contained within each file including careplans, conditions, encounters and observations. The user is also given a plot of patient observations, of which the observation is selected by the user. For presentation purposes, the example interface allows selection of the BMI, insulin resistance and H1bAc levels however it can be populated with all observation measurements.

3. Patient Genomic Data

Each patient ID from the Synthea clinical data was assigned a synthetic gene expression profile. The gene expression profiles were then assigned to replicate the patterns that would be associated with the control and diabetes patients. The top genes associated with type 2 diabetes were extracted and significant expression levels were given to diabetes patients. Other genes "measured" were given a non significant value for the control samples. For example, an

individual who has evident insulin resistant levels was given a NAT1 (insulin resistance associated) expression level that would be significantly higher to mimic the associated.

4. Cohort Clinical Data

The cohort data was used to identify subgroups within the samples. These subgroups include the number of disease cases, the ethnicity breakdown, the group age and changes in BMI throughout time. These changes could be used to target group treatment or identify new patterns. For example, by narrowing the group into individuals who of Irish descent, it may be possible to identify new gene expression patterns that would be not be previously identifiable in the mixed groups. Also knowing the number of diseases can be used for technical issues like batch control. The changes over time can also be identified to understand if environmental factors are at play.

Conclusion

Using the longitudinal data that is captured within EHRs, new patterns can be identified in disease progression that are not possible by individual record inspection. The graphical representation of the data results in faster diagnosis and improved healthcare by allowing the development of predictive models and risk estimation. By visualizing the healthcare data using dashboards, patient stratification is accelerated and the trends in patient health can be used to identify the right time to intervene. However, the lack of a platform to organise the data into a way that it can be visualised, has resulted in slow adoption of a cohort based approach. Providing a platform that requires little expertise and quick visualisation of these patterns, will provide sufficient structure for the efficient inspection of the data. The application has been verified by using a GEO sample to ensure that gene expression data can be integrated with genomic.

Several limitations must be addressed within this application. First of all, the data used is synthetic. Synthea indicate that although they are confident in the ability of the app as a framework for development, it is important to consider it only as sandbox data for development as opposed to decision making data. So while it fulfuls that function here, the results for the clinical patterns should not be relied on. It should also be noted that there are issues with the categorising system: for example some entries for the disease are *History of myocardial infarction (situation)* and *myocardial infarction* which are the same but fall under different Categories. Synthea also contains discrepancies with demographic data that have been noted (ref). Synthea developers have addressed this and are continuously improving the likelyness to real data through their GitHub account (link).

Synthea however did prove very useful for the acquisition of clinical data - it improved efficiency by allowing instant usage and data manipulation to model the cohort effectively and eliminating time associated with data cleaning given that clinical data is so commonly ‘dirty’. Efficiency is also improved by the elimination of issues associated with privacy including de-personalising the data and requesting new data privileges.

Discussion

This framework allows interoperability among departments and the use of clinical and genomic data for its relevant function. The use of as standardised format such as the FHIR7 means that the method can be applied to other forms of the data. However, the need to

Figure~2\vphantom{\ref{fig:02}} Talk about a second use of I2EHR. Talk about how it was implemented through the treatment of type 2 diabetes patients \citealp{Boffelli03} might want to know about

or new

The graphical data are a systemic way to analyse clinical observations and treatments and influence the focus for genomic research.

This approach intends to give clinical researchers the insight required to improve diabetes treatment and help them to select the right genomic analysis that need to be performed to gain useful knowledge on the molecular background to the disease.

Limitations of microarrays: limitations include cross-hybridisation induced high background limitations

To discuss

- Motivation
- The motivation for the study was to develop a framework that intergrated both clinical and genomic data; this framework was generated given the lack of such as system and the expected functionality that would result from it. The system has proven that it is possible to extract data and use it for this function.
- Set up Synthea
- The clincial data generation also had issues regarding the control samples. As a synthethic patient generator, Synthea doesn't allow the generation of a sample of patients that are healthy but instead gives only access to patients that have been disease diagnosed. This proved a challenge for demonstrating the difference in expression vs the control sample. R
- Set up GEO

- The extraction of data using GEO also proved to be more challenging that expected for finding a suitable sample of gene expression data. GEO datasets that contained a large sample of gene expression data could not be found that exceeded 12 patients. To address this, values were generated for the expression level that would be expected for a diabetes type 2 patient. This meant that the entire analysis would be a simulation of data. This decision was made following searching ArrayExpress and also finding no larger usable samples.

- Reason for analysis types
-
- Set up for the