Abstract

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 which predates the widespread utilisation of cohort analysis. 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.

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. I also used a cohort of diabetes patients that were artificial and generated using a synthetic patient population generator called Synthea. Further, the integrated data could be used to create predictive models depending on the patients expression data and clinical data. 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.

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

Results

Discussion

Talk about what the framework is and how easy it is to use the I2EHR is.
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 text text text text
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.

Synthea proved very useful for the acquisition of clinical data; Synthea improved efficiency by allowing instant usage and data manipulation to model the cohort effectively and eliminating waiting time. Efficiency is also improved by the elimination of issues associated with privacy including de-personalising the data and requesting new data privileges.
Issue with Synthea is the 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).

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