Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 27 June 2019
Data and text mining

I2EHR: Interactive Integrated Electronic Health Records

Shane Crinion *, Pilib Ó Broin

[1] School of Mathematics, Statistics & Applied Mathematics, National University of Ireland, Galway, H91 H3CY

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Electronic health records contain dense, longitudinal data that cannot be used to full potential due to a lack of standardisation and infrastructure. A framework that allows integrative analysis of clinical and genomic data is likely to improve healthcare delivery and predictive analytics.
Summary: We present the Interactive Integrated Electronic Health Record (I2EHR), a dashboard built using R/Shiny allowing integrative patient and cohort analysis across operating systems. The interface allows the user to view patient or cohort level and generate visualisations based on user-selected measurements. This model, generated using realistic synthetic clinical data, is then integrated with gene expression profiles to develop a realistic and longitudinal model of disease pathology. Lastly, the model uses the newly identified genotype-phenotype associations to perform predictive modelling.
Availability: I2EHR can be obtained from GitHub: `github.com/shanecrinion/I2EHR`. The subdirectory "I2EHR_APP" contains the application script and clinical data files. The genomic data is downloaded within the script from the Gene Expression Omnibus and is accessible under the accession number GSE46097.
Contact: s.crinion1@nuigalway.ie
Supplementary information: Supplementary data is available from the Supplementary Information file on the I2EHR online repository.

1 Introduction

The widespread implementation of the electronic health record (EHR) substantially improves the quality and efficiency of healthcare service delivery [1]. By using EHR systems, an extensive and longitudinal profile of the patient's historical recordings can be built at the point of care. Substantial improvements to healthcare delivery and research can occur following integration of EHR systems; the use of digital records in healthcare has significantly reduced errors and increases the amount of usable data for clinical research [2]. Advanced EHR systems that enable clinical decision support (CDS) are used for decision making, forecasting patient outcomes and modifying treatment promptly to improve delivery of diagnosis and drug prescription [3]. Given the exponentially growing size of clinical and genomic data available, the EHR provides a storage method beyond human capacity [2]. Data stored in EHRs including vital signs, billing data, laboratory results or genomic data in the most advanced systems [4].

However, the most useful function with EHR data may be the potential for text mining with patient data. Longitudinal, dense patient data stored within EHRs is especially useful for understanding disease epidemiology and progression [5]. Providing access to high quality lifetime patient data is extremely useful for improving treatment of complex, heterogeneous diseases [6] including type 2 diabetes (T2D) and cancer.

The user interface (UI) of EHRs is commonly designed poorly and many lack the functionality to view numerous clinical recordings simultaneously [7]. At patient level, a timeline or other visual representation may improve understanding of changes in patient health over time. Risk factors that signal disease progression for example, high BMI in T2D or high blood pressure in cardiovascular disease (CVD), may show up that are not obvious with static data.

Time-based visualisations can also identify novel risk factors that may not be obvious throughout disease progression. Identifying new links between clinical measurements and patient outcome could be used to improve prognosis prediction or CDS functionality such as drug selection based on disease severity. However, the potential for a clinician to find new links or assess patient health recordings is greatly limited by the

need to manually examine changes. Improving the digital readability of clinical data increases the data research potential as large cohorts can be built and studied. Researchers can use structured clinical data to study changes in the cohort's clinical recordings due to deterioration in health or drug prescription. For example, a cohort of opioid addicts accessed through EHR data would be useful for identifying changes due to drug use or warning signs associated with death [5].

In order for clinical data to be integrative, numerous challenges need to be addressed including the lack of standardisation in data format. The development of integrative tools in healthcare has been limited by the wide variety of healthcare data available, causing reduced ability to process patient information and a lack of software for data analysis [1]. Concerns regarding privacy of sensitive patient data further exasperates the lack of medical software [8]. The use of personal patient information requires that prior consent and adherence to legal and technical protections are followed which can result in delayed or restricted processing of patient data.

Approaches to improve health informatics have addressed the data entry format, the interactivity of the data format and the availability of test data. HL7 Fast Healthcare Interoperability Resources (FHIR) [9] is a clinical data standard that provides a framework to improve data entry and subsequently increase the interoperability and potential for data exchange between EHRs. ShinyFHIR is an application that allows interaction with FHIR-based data and demonstrates how standardised data can be used the enhance clinical analysis [9]. The ShinyFHIR approach demonstrates the suitability of the FHIR approach to work for numerous servers and across platforms for clinical data analysis using the *R* and *Shiny* packages. Synthea is an appraoch that address the lack of available clinical data by producing synthetic but realistic EHR data that can be used for medical software development [10]. Usage of diagnosis codes have also improved standardisation and allowed the improved exchange of notes between departments. Using diagnostic codes also improves computational readability and subsequently facilitates machine learning approaches for improved prediction of disease progression and treatment plan selection.

The integration of genomic data to EHRs (such as in biobanks) have been shown have huge innovative potential and potential to improve clinical research [4]. However, a lack standardisation of clinical data format also limits the potential for genomic data integration. Many limitations exist which prevent the integration of genomic data into genomic health records; one example is the lack of training or knowledge in genomic analysis result can result in poor interpretation of genetic variations and significance levels in disease diagnosis [11].

To address current challenges associated with the integration of clinical and genomic data, we designed a model to integrate data and provide visual and timeline representations of patient recordings. As outlined in Figure 1, the aims of this project include (a) to develop a dashboard to integrate clinical and genomic data (b) create simple predictive analytics using the patient and clinical data and (c) use GEO to integrate gene expression profiles 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 applied to many hospital EHRs to provide useful statistical results for genomic research. The tool requires little expertise to use and 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 at patient or cohort level and further improve the utilisation of big data in genomics.

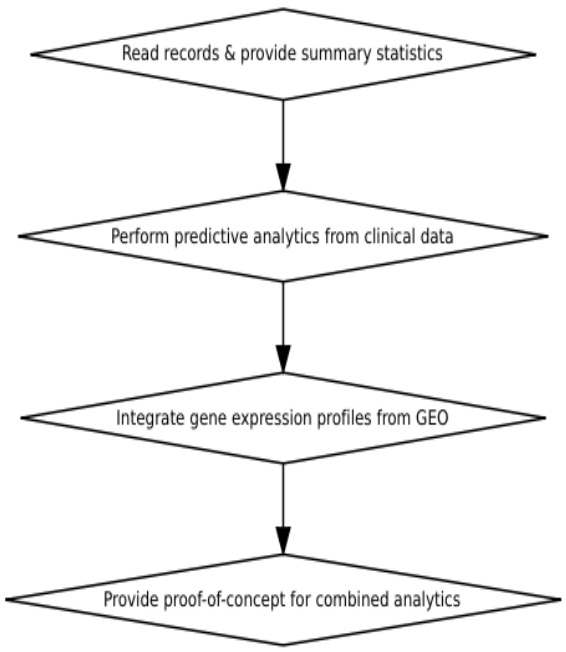


Fig. 1. Analysis flowchart: Clinical data is read, analysed and integrated with genomic data. Clinical observations are then integrated with genomic expression values to visualise subgroup variation.

2 Materials and Methods

In order to develop the integrative application, the clinical data came from Synthea [10], a package used to generate synthetic, realistic EHR data. Synthea was developed due to the lack of clinical data available for use in software development and other research or education purposes without any concerns for legal, privacy or security issues [10]. The package also follows the standardised FHIR HL7 format and provides a source of 'clean' structured data which contains no discrepancies due to human error. Synthea is built with the top-down approach named Publicly Available Data Approach to the Realistic Synthetic EHR (PADASER) which is a framework to generate EHR data coded in the HL7 FHIR format. The PADASER framework uses publicly available datasets to populate synthetic EHR including health incidence statistics, clinical practise guidelines (CPGs) and Protocols and Medical Coding Dictionaries. The approach has a core principal of privacy and avoids any risk of patient re-identification as reported in previous studies [12].

At the core level, Synthea consists of various generic modules to model the 10 most common reasons for patient visits in the United States according to the Global Burden of Disease [10]. The patients generated were from the 'Cardiovascular Disease' module which consists of coronary heart disease, myocardial infarction, cardiac arrest, stroke and atrial fibrillation. The 'wellness' module was also used which contains clinical measurements such as body mass index (BMI), diastolic blood pressure and pain severity which can be useful for visualising change over time.

The original population generated consisted of 1000 patients originating from Boston, MA. The output consists of eleven CSV files

Observation (units)	Data Source	Normal Level	CVD Level
BMI	Clinical	<30	>30
Diabetes (mmol/L)	Genomic	>5.6	>=7
Diastolic BP (mm Hg)	Clinical	<80	>80
Systolic BP (mm Hg)	Clinical	<120	>120

containing various types of clinical data namely allergies, careplans, conditions, encounters, imaging studies, immunizations, medications, observations, organizations, procedures and providers. Each file contains clinical information for patients using their patient ID number. The patients file contains demographic information such as patient ID, first name, last name, address, gender, ethnicity and race (figure 3). The conditions file includes the associated condition for the patient (for example coronary heart disease for patient 1e48d3da-d4f0-4680-8106-2304c9d1426e). The encounters file records each patient encounter, all of which are under "Encounter for check up (procedure)" or "Emergency Encounter". The observations file contains every clinical measurement for the patient generated from the wellness module such as BMI and diastolic blood pressure. The procedures indicates the treatment provided for the patient, for example "electrical cardioversion" and "Catheter ablation of tissue of heart". The population had no generated information for allergies, careplans, imaging studies or providers. The data contained within each file was used to build a longitudinal patient profile and to differentiate subgroups within the cohort.

Measurements extracted from the observations file were selected to model the time dependant variation of patient

Clinical measurements from the observations file were selected to model the variation expected for a cardiovascular disease patient. BMI, diastolic blood pressure and systolic blood pressure were selected as the outcomes of interest for subgroup selection. A BMI measurement of above 30 is considered high and is expected for an individual with a cardiovascular disease [13]. Similarly, the diastolic and systolic blood pressure measurements are expectedly elevated in a cardiovascular disease patient and measured above 80 and 120 mm Hg respectively [13]. These measurements are recorded in table 1.

The aims of this project include integration of clinical data with gene expression profiles to allow combined clinical / molecular analytics. Gene expression profiles were sourced from the Gene Expression Omnibus data repository which contains many publicly genomic data sets including expression by array sets. *"Expression data of Participants of Ornish intervention and Control group"* was selected, a study that aimed to examine the impact of a CVD risk reduction program on peripheral blood gene expression profiles in 63 Caucasian participants and 63 matched controls following rigorous lifestyle modification and identify regulatory pathways associated with cardiovascular health. Entry criteria for the at-risk participants included diagnosis of coronary artery disease (CAD) and diagnosis of diabetes. Controls were matched based on age, sex and disease status. The number of participants with diabetes and CAD was matched with controls and consisted of 10 diabetes and 23 CAD samples each.

The expression data consists of 379 CEL files with 3 samples for each individual taken at baseline, 3 months and 1 year. Genotyping was performed using [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array. The data was background corrected using robust multi-array average (RMA) and log2 transformed prior to integration. The data was extracted GEOquery, a Bioconductor package that facilitates import of microarray data directly from R using the GEO accession number [14]. The command `getGEO("GSE46097")` imports the data as an ExpressionSet object which combines the sample phenotypic data and experimental chip and technology data used for the experiment. The function `exprs` is used to extract the per probe expression levels.

To integrate the clinical and genomic data, each sample from the generated Synthea data was assigned a GEO accession corresponding with a case or control sample. Each sample within the genomic data was subsequently assigned a patient ID. Given that Synthea does not allow control sample generation, the conditions file was manipulated to label the control patient's condition accordingly. The clinical observations values for BMI, diastolic blood pressure and systolic blood pressure were categorised based on the thresholds within table one. For the participant vs. control relative log expresssion plot (RLE), the associated control was not labelled as such. A control was selected by subsetting to samples that was within five years of age with matching diabetes status, CAD status and gender. Additionally, the genomic data was subset to baseline and 1 year samples only for the t-test step within the patient level clinical analysis.

The development of this dashboard was made possible using the packages R and shiny R version 3.6.0. The user can obtain the app from `github.com/shanecrinion/I2EHR/I2EHR_APP`. The front-end packages used for app development included the following: shinydashboard and shinywidgets. Visualisations were generated using the ggplot2 and plotly packages. Associated packages include: ggridges, lattice, viridis, DiagrammeR, gplots, ggplot2, geneplotter, RColorBrewer and pheatmap. The integration and analysis of genomic data was made possible using multiple packages from Bioconductor which include GEOquery, oligo, arrayQualityMetrics and the analytics and statistics packages limma, topGO, ReactomePA and clusterProfiler.

3 Results

I2EHR contains 6 major components that were deemed most relevant for the integrated analysis of the data: (1) patient level data query (2) patient level genomic analysis (3) cohort level observations plot (4) principal component analysis (5) subgroup expression analysis (6) microarray analysis

3.1 Patient level data query

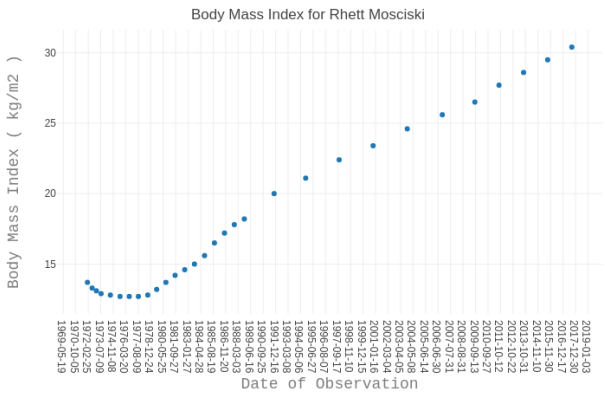
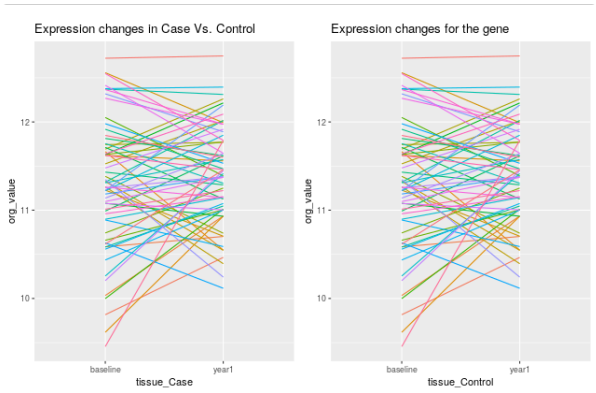


Fig. 2. Observation plot: The x-axis indicates the date of observation and y-axis indicates the measurement. A drop down menu allows selection of the observation to observe over time.

To query a patient, select the "Patient Analysis" tab and enter the patient ID number in the search bar. A table including the patient ID, full name, gender, ethnicity and race is generated. To query an observation, select a type from the drop down such as "Body Mass Index". The patient observation type and name populates a plot. The plot is generated using

plotly package which allows the user interact by zoom, pan or hover to display corresponding information for the selected data point. The name, description, units and dates are extracted to populate a unique plot for the patient/observation. Under the heading ”Additional Information”, a table can be generated containing all other patient recordings from of the CSV files.

3.2 Patient level gene expression variation




```
        Paired t-test

data:  RPL35_year1 and RPL35_baseline
t = 1.322, df = 62, p-value = 0.191
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05524793  0.27102825
sample estimates:
mean of the differences
              0.1078902
```

Fig. 3. Expression changes: each plotted line represents an individuals variation in gene expression between baseline and 1 year assessment.

As seen in figure 3, the dashboard allows a visualisation of patient level gene expression variation. A table of the plot contains all genes and probe information; this can be used to select a probe and view gene expression changes in the case and control. Additionally, a t-test is generated below indicating the significance of the case and control samples variation.

3.3 Cohort level barcharts

Cohort level barcharts are available within the ‘Cohort’ section on the dashboard sidebar. As seen in Figure 4, barcharts are used to visualise the dispersal of clinical entries for the cohort. Buttons allow the user to select between viewing frequency of recordings for immunizations, medications, procedures and conditions. This can be used to understand the most common diseases or treatments being applied for the cohort.

3.4 Principal component analysis

Principal component analysis (PCA) is used to identify possibly correlated variables and visualise genetic distance. The PCA plot of Figure 5 demonstrates the integration of clinical and genomic variables. The user selects the colour fill of the PCA from the three clinical data measurements (BMI, SBP and DBP) and two genomic data measurements (diabetes and CAD status).

3.5 Gene expression density plots

The gene expression density plots

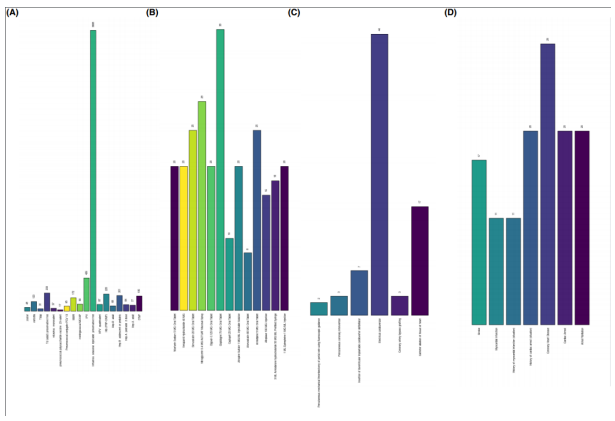


Fig. 4. Barcharts: Plots indicate the frequency for each of (A) immunizations, (B) medications, (C) procedures and (D) conditions at cohort level.

3.6 Microarray analysis

Given the integrative design of the application, the microarray analysis steps are also contained within the app to provide an all-in-one integrative research application. This allows the user to upload their CEL file and perform each step of the analysis. This can be used by the clinician to examine the quality of the data and ensure that the data of high quality before generating any results using it. The layout of the microarray analysis as seen in Figure 4 follows standard procedure and displays a tab for each of quality control, normalisation, differential expression analysis and biological interpretation.

The quality control displays a number of visualisations which can be used for identifying any variation at the probe level and ensure that no sample reading is considerably different to others. The quality control includes a between array comparison, an array intensity distribution, variance, standard deviation from the mean and individual array quality. The normalisation function will indicate the array information to identify whether the values need to be normalised. Normalisation of data depends on the type of array, design of the experiment and assumptions regarding the microarray expression. The normalisation method used is dependant on the sequencing - Affymatrix data is normalised using the Robust Multi-Array Average (RMA) method and the oligo package. The next tab, differential expression analysis, is used to identify the variation due to the risk factor. The log2 fold change and multiple testing correction is performed using limma‘ on Expression Atlas data. The interpretation is through heatmaps, functional annotation and network analysis. This can then be used to identify the sample relationships and the genes that are associated with the disease in question.

4 Discussion

The motivation for the study was to develop a framework that integrated both clinical and genomic data; this framework was generated due to lack of a similar system and the potential benefits for research. I2EHR demonstrates data collected in FHIR format can be integrated for exploratory purposes and integrated with genomic data on a single and easy-to-use usable platform.

Visualisations of longitudinal clinical data proved useful for identifying changes over time associated with T2D such as BMI variation. I2EHR facilitates further identification of new risk factors for disease progression by plotting observations on a patient or cohort basis. Further to this, risk factors can be used for predictive analytics. By visualizing the healthcare data using dashboards, patient stratification is accelerated

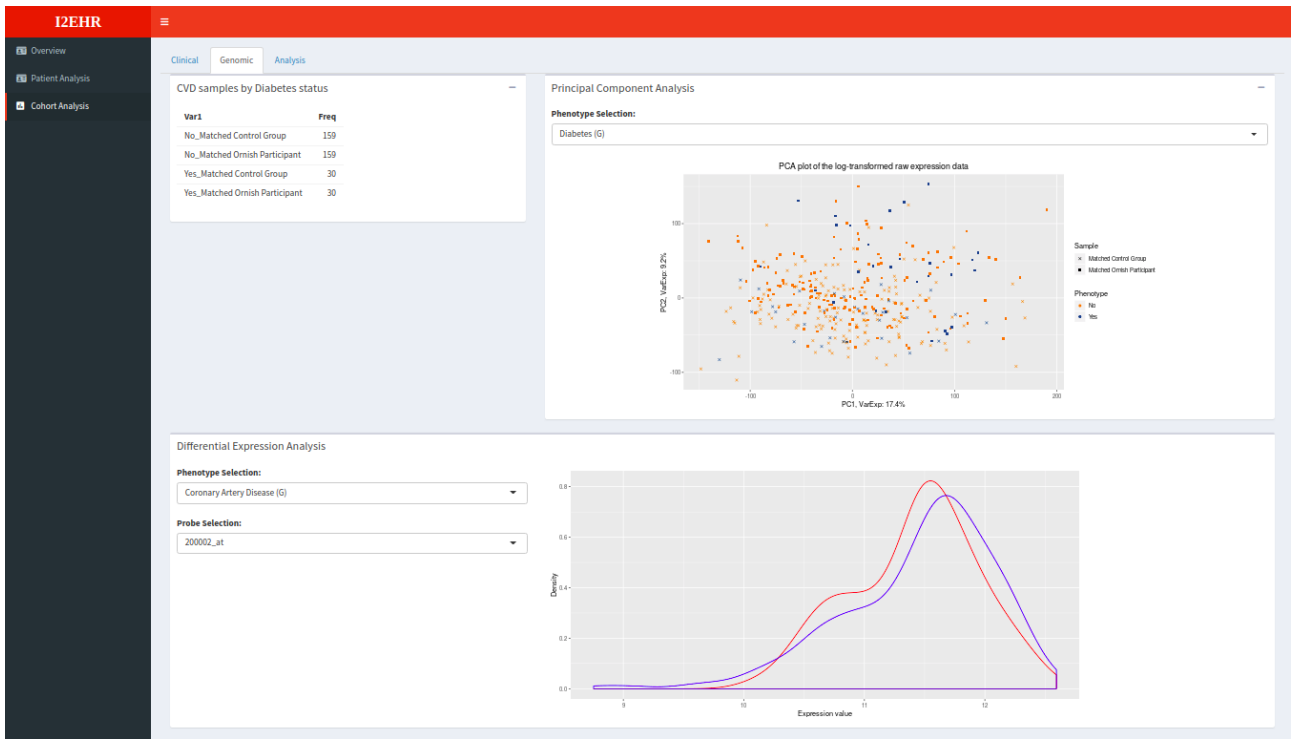


Fig. 5. Dashboard Layout: I2EHR UI including the PCA and density plots allowing subgroup selection. This can be used to identify subgroup clustering or gene expression variation.

densityplots.png

Fig. 6. Density plots: The generated plots indicate the gene expression variation found between the participant (blue) and control (red).

and the trends in patient health can be used to identify the right time to intervene. The graphical representation of the data results in faster diagnosis and improved healthcare by allowing the development of predictive models and risk estimation with an easy, accessible interface.

Issues regarding standardisation have caused difficultly in the adoption of plots and other visualisation aids. Heterogenous, complex data within current EHRs due to manual entry of data and a lack of structure standards has further reduced the ability to intergrate genomic data. In the absence of standardised methods such as the FHIR entry approach, it is virtually impossible to associate the observation entry with the gene expression. This platform provides an example of the capabilities of clinical and genomic data integration that can be applied to other data formats.

Normally EHRs platforms do not support simultaneous view of the clinical data entries with genomic data. I2EHR allows the user to view their data entries in table format or view the clinical observations using the plotly generated plots. A core element for widespread implementation is ease of use - presented is a dashboard interface allowing easy access and interactivity for the user. The user can view clinical observations, genomic expression and quality control in one place. This allows the user to ensure that the data is suitable for use before taking further clinical action. Providing a platform that requires little expertise and quick visualisation of these patterns, will provide sufficient structure for the efficient inspection of the data.

Several limitations must be addressed for Synthea. First of all, the data used is synthetic. Synthea developers indicate that the data is useful for IT developments and other purposes however the clinical observations and statistics should be considered with discretion. Synthea data is used as a framework for development and it is important to consider it only as sandbox data for development. No clinical decision making should be performed using the data and use should be for proof of purpose only. 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 categorisation. Synthea also contains discrepancies with demographic data statistics that have been noted in the issues section of their GitHub account and in other reviews. Synthea developers are continuously improving the data to stay current.

Another caveat using Synthea occurs when generating control samples. As all samples available are disease patients, each individual has a diagnosis of some sort. Synthea does not allow the generation of a sample of patients that are healthy but instead gives only access to patients that have been disease diagnosed. This may cause discrepancies when comparing differential expression identified in this dataset with real cases.

Despite the caveats, Synthea has proved highly useful for the objectives of the project. Synthea improved efficiency by allowing instant data usage and manipulation for effective modelling of a T2D cohort. This has also eliminated time associated with data cleaning given that clinical data is often dirty due to errors or non-structured data entry. Efficiency is also improved by the elimination of issues associated with privacy including de-personalising the data and requesting new data privileges.

This framework developed with the help of Synthea allows interoperability among departments and the use of clinical and genomic data for its relevant function. The use of standardised format such as the FHIR means that the dashboard can be used with other sources of data in the same format. Further emphasis on data standardisation is a core component for utilisation of I2EHR and the wider capabilities of integrated clinical and genomic research.

The use of integrative data is particularly useful for diseases such as diabetes type 2. A disease like diabetes shows many characteristic risk factors such as hypertension, high insulin resistance and glycated hemoglobin. 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. The development of approaches such as the Health Information Technology for Economic and Clinical Health (HITECH) Act have resulted in more meaningful use of data related to diabetes patients [15]. Given that diabetes is often manageable through diet control, this is especially useful for prevention of disease progression. The use of visualisations makes the analysis of diabetes related measurements more accessible and easy to understand.

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 with diagnosis of T2D and no other disease. The alternative data repository ArrayExpress [16] also showed no results for larger datasets. To address this, values were generated for an expected expression level for a T2D patient. Gene expression profiling is complex and this complexity was a core reason for using simulated data. As mentioned above, no large T2D samples were available online and half of the sample had colon cancer in the largest set. Given the ambiguity that would be expected with this dataset, it was decided to use simulated values to avoid confusion and continue with an accurate model of the expression that expected for a diabetes patient.

For application development, Shiny proved especially useful for the development of useful software for use in analysis of clinical and genomic data. By using Shiny, other bioinformaticians will be able to use their expertise in clinical settings without the need for advanced software development skills. Without experience, interactive plots could be made using plotly which allowed the user to focus on certain periods in the patients lifetime. The app also allowed overlay of the clinical and genomic data to view the gene expression pattern in relation to the clinical observations and potentially identify new associated between gene expression and clinical observations.

In the case of genomic data, it allows the patient to use raw data and perform each step of the microarray analysis in a self contained tabs including the quality control, normalisation, differential expression analysis. Following this, numerous graphs are generated which can be used for the clustering analysis, gene set enrichment and the pathway or network analysis. By containing each of these steps within the app, it allows the clinician to correct the data themselves and removes the need to contact a bioinformatician in order to clean the data and extract the useful information.

To conclude, I2EHR provides a platform that can be used across platforms for the integrated analysis of clinical and genomic data. Further developments will be possible given greater adoption of a standardised data format in EHRs.

Acknowledgements

I would like to thank Dr. Pilib Ó Broin for the helpful advice and guidance throughout the project. I also would like to thank the developers of R, Shiny and Synthea for their applications that allowed this development to happen.

Conflict of interest

No conflict of interest is declared.

References

[1]R.S. Evans. Electronic health records: Then, now, and in the future. *Yearbook of medical informatics*, pages S48–S61, 2016. cited By 14.

[2]Marc S. Williams. The genomic health record: Current status and vision for the future. In Reed E. Pyeritz, Bruce R. Korf, and Wayne W. Grody, editors, *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics (Seventh Edition)*, pages 315–325. Academic Press, January 2019.

[3]Brian Rothman, Joan C. Leonard, and Michael M. Vigoda. Future of electronic health records: Implications for decision support. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 79(6):757–768, 2012.

[4]Joshua C. Denny. Mining electronic health records in the genomics era. *PLOS Computational Biology*, 8(12):e1002823, December 2012.

[5]Joan A. Casey, Brian S. Schwartz, Walter F. Stewart, and Nancy E. Adler. Using electronic health records for population health research: A review of methods and applications. *Annual Review of Public Health*, 37(1):61–81, 2016. PMID: 26667605.

[6]A.C. Faria-Campos, L.A. Hanke, P.H.S. Batista, V. Garcia, and S.V.A. Campos. An innovative electronic health record system for rare and complex diseases. *BMC Bioinformatics*, 16(19), 2015. cited By 1.

[7]Nir Menachemi and Taleah H. Collum. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, 4:47–55, May 2011.

[8]M. K. Ross, W. Wei, and L. Ohno-Machado. "big data" and the electronic health record. *Yearbook of medical informatics*, 9(1):97–104, August 2014.

[9]Na Hong, Naresh Prodduturi, Chen Wang, and Guoqian Jiang. Shiny FHIR: An integrated framework leveraging Shiny R and HL7 FHIR to empower standards-based clinical data applications. *Studies in health technology and informatics*, 245:868–872, 2017.

[10]J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc*, Aug 2017.

[11]Rebecca Fein. Innovate or die!: Genomic data and the electronic health record (ehr). *Applied & translational genomics*, 3(4):130–131, 2014.

[12]Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6:e28071, 2011.

[13]Darrell L Ellsworth, Daniel T Croft Jr, Jamie Weyandt, Lori A Sturtz, Heather L Blackburn, Amy Burke, Mary Jane Haberkorn, Fionnuala A McDyer, Gera L Jellema, Ryan van Laar, et al. Intensive cardiovascular risk reduction induces sustainable changes in expression of genes and pathways important to vascular function. *Circulation: Cardiovascular Genetics*, 7(2):151–160, 2014.

[14]Sean Davis and Paul S Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.

[15]V. Patel, M.E. Reed, and R.W. Grant. Electronic health records and the evolution of diabetes care: A narrative review. *Journal of Diabetes Science and Technology*, 9(3):676–680, 2015. cited By 2.

[16]Awais Athar, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, Nuno A Fonseca, Robert Petryszak, Irene Papatheodorou, et al. Arrayexpress update–from bulk to single-cell expression data. *Nucleic acids research*, 47(D1):D711–D715, 2018.