

Data and text mining

Interactive Integrated Electronic Health Records

Shane Crinion *, Pilib Ó Broin

¹School 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 and interactive analysis of clinical and genomic data using the HL7 Fast Healthcare Interoperability Resources standard framework is likely to improve clinical delivery of care.

Summary: We present the Interactive Integrated Electronic Health Record (I2EHR), a dashboard built using R/Shiny allowing integrative patient and cohort analysis across platforms. The interface allows the user to view patient and cohort level data interactively 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. This includes the instructions for how to install required packages and use the package. The GitHub account also contains the sandbox synthesised patient data for type 2 diabetes patients to explore the application usability.

Contact: s.crinion1@nuigalway.ie

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 data, an extensive and longitudinal profile of the patient's historical recordings can be built at the point of care. Healthcare delivery and research improves accompany EHR adoption; digital storage has reduced error rate from 64% to 0% ($p = 0.0001$) in the Natural Hip Fracture Database [2], greatly increasing usefulness for research and auditing. Both the clinician and patient also benefit from clinical decision support (CDS) which has proved useful for decision making, forecasting patient outcomes and modifying treatment promptly to prevent incorrect prescription or diagnosis [3]. Given the exponentially growing size of clinical and genomic data available, the EHR also provides a storage mode method beyond human capacity [2].

However, arguably the most useful function possible using EHR data is patient data text mining. Longitudinal, dense patient data stored within EHRs is especially useful for understanding disease epidemiology and progression [4]. Providing access to high quality lifetime patient data

is extremely useful for improving treatment for complex, heterogeneous diseases [5] including type 2 diabetes and cancer.

The current user interface (UI) for EHRs lacks the ability to view numerous recordings simultaneously. On a patient level, the use of a timeline or other visual representation can improve understanding changes in patient health over time. This would allow the clinician to view risk factors that signal disease progression such as high BMI in type 2 diabetes. The lack of this time-based functionality also limits the ability to identify new disease risk factors warning signs for disease progression. By viewing the patient's wellness measurements all together, Currently, time consuming manual data review of the patient information is required and limits potential for text mining.

New links between clinical measurements and outcome may be used for prediction of prognosis or clinical decision support functions eg. drug selection based on disease severity. The access to cohort of patients also has great improvements for research including access to large patient numbers without recruiting and testing the exposure to an unethical risk factor [4].

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. The lack

of standardisation in turn has resulted in health technology lagging [1] and inefficient sandbox data to develop software. Concerns regarding privacy of sensitivity patient data exasperates the lack of developments within healthcare [6]. FHIR provides patterns and best practise to use for detectable and standardised data entry in health records

Approaches to improve health informatics have targeted the data entry format, the interactivity of the data format and the availability of test data. The clinical data standard HL7 Fast Healthcare Interoperability Resources (FHIR) [7] provides a framework to improve interoperability and data exchange in EHRs. To address the potential benefits for analysis, approaches like ShinyFHIR have integrated FHIR data into R for patient and cohort analytics [7]. 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. Other approaches such as Synthea address the lack of available clinical data by producing synthetic but realistic EHR data to be used for health informatics. Use of clinical diagnosis codes improved this further and allowed the improved categorisation and population of notes between departments. This improves computational readability and results in improved prediction of disease progression and treatment plan selection.

The extension of genomic data in EHRs in approaches such as biobanks have demonstrated the potential for pattern identification possible with the integration of clinical and genomic data. The lack of standardisation limits the ability to integrate data. Many limitations exist which prevent the integration of genomic data into the genomic health record; one example is a lack of adequate training which can cause issues regarding interpretation of the genomic data. Results may not be interpretable with bioinformatics knowledge; conversely genetic risk variants may be considered sufficient for diagnosis if genomic data is analysed without appropriate discretion. (Rebecca Fein, 2014)

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, I aim to improve the accessibility of combined analytics and a patient and cohort level and further improve the utilisation of longitudinal patient data.

2 Materials and Methods

In order to develop the integrative application, the clinical data came from Synthea [8], 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. The package also follows the standardised FHIR HL7 format and provides a source of 'clean' structured data without data entry related discrepancies. 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) Protocols and Medical Coding Dictionaries. The approach has a core principal of privacy and avoids any risk of patient re-identification as reported [9] in previous studies.

At the core level, Synthea consists of various modules to model the 10 most common reasons for patient visits (with diabetes being the 3rd most common) according to the Global Burden of Disease for the United States [8]. The population generated here consisted of 1000 individuals originating from Massachusetts, MA. The patients files contains demographic information such as first name, last name, address and gender. The selection criteria specified that patients suffered from a metabolic disorder. Diabetes is accessed using 2 of the models available, namely namely metabolic syndrome disease and metabolic syndrome care. These 2 modules were used to include diabetes specific measurements such as glycated hemoglobin (Hb1Ac) and insulin resistance measurements. Other general measurements were generated using the 'wellness' module including height and weight. These measurements were compared and contrast the models generated for diabetes and control patients. These measurements, along with other wellness measurements, are accessible from the observations file.

Insulin resistance and HbA1c levels were selected as the outcome-of-interest that would be used to identify subgroups of the disease. 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

Source: American Family Physician . 1/15/2016, Vol. 93 Issue 2, p103-109. 7p. 5 Charts. Author(s): PIPPITT, KARLY; LI, MARLANA; GURGULE, HOLLY E. Diabetes Mellitus: Screening and Diagnosis.

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 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).

In addition to recording HbA1c and insulin resistance measurements, the relationship between drug prescription and diabetes was desired. The commonly prescribed diabetes drug Simvastatin (20 MG Oral Tablet) was selected for the simulation. This was one of the most commonly prescribed medications found in the Synthea patients.

The aim of this project included integrating clinical data with genomic data as proof of purpose for integration. Genomic data sources which were included to display integrative analysis including expression profiling by microarray, and synthetic expression profiles. Gene expression profiles were sourced from the Gene Expression Omnibus data repository. GEO Profiles was searched with the search term "type 2 diabetes" AND ("50" [n_samples] : "1000"[nsamples]) and filtered to only human samples. Of the top 20 results sorted by subgroup effect, 12 results came from one series - GSE25462. Montemanto et. al indicate that CAT, FTO and UCP1 are all associated with type 2 diabetes. Searching the terms type 2 diabetes directs to the same series: GSE25462. This dataset titled Type 2 diabetic and insulin-resistant but normoglycemic cohorts: skeletal muscle was designed to identify the relationship between hereditary insulin resistance and diabetes and consists of 10 diabetic patients, 25 subjects with a family history of type 2 diabetes (one or both parents), and 15 subjects with no family history of type 2 diabetes. 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. This application uses this sample for the purpose of presentation however recognises that the sample

size is not sufficiently powerful enough to accurately interpret the data results.

As previously mentioned, the drug Simvastatin (20 MG Oral Tablet) was selected its effect on diabetes patients. Searching for the term "type 2 diabetes "Simvastatin" in GEO profiles, indicated 30 results for gene APOC3. APOC3 was therefore selected as a candidate gene for modelling the relationship between drug prescription and gene expression variation. The top result measured APOC3 levels in the dataset GSE25462 again. A link between mitochondrial oxidative homeostasis and predisposition to diabetic vascular complications has been previously identified (Alberto Montesanto, 2018). Plasma APOC3 associates with almost half of HDL (Zvintzou E, 2017) while HDL-associated apoA-I and lysosphingolipids protect against ATP reduction resulting from reverse cholesterol transport in mitochondria (CR White, 2017).

Considering the small sample size of the first set, another sample of a larger size was considered - GSE115313. This dataset titled Transcriptomics analysis of paired tumor and normal mucosa samples in a cohort of patients with colon cancer, with and without T2DM. This study aimed to identify the relationship between type 2 diabetes and colon cancer. The samples included "2 types of samples from 42 patients with colon cancer: i) tumor samples and ii) normal colonic mucosa. The cohort is composed by 23 non-diabetic patients and 19 diabetics". This study was selected by searching type 2 diabetes and sorting by the highest number of patients with type 2 diabetes. The aim of selecting this dataset was to use a larger sample set and subsequently find more significant subgroup effects. However, considering that the samples were half colon cancer samples, it was decided no to use these due to the ambiguity as to whether gene expression variation was due to cancer or diabetes.

Finally, given that the target disease group for the analysis was type 2 diabetes, it was decided that the best way to perform the analysis would be by using expression values that would mimic those expected for a diabetes patient. To do this, a matrix was generated (below) that contained expression values expected for diabetes patients. This expression average was first obtained from the insulin resistance related dataset above. Following this, a list of all the most highly associated genes was obtained from X. These genes were given an expression value that was concordant with that expected for a diabetes patient. This decision was made given that there were no alternative datasets publicly available that would be large enough for the desired effect and meant that the entire application was built using a synthetic simulation.

```
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])
```

The development of this dashboard was made possible using the packages R and shiny R version 3.6.0 Planting of a Tree. The user can obtain the app from https://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. To run the app type runGitHub(shanecrinion, I2EHR, I2EHR_APP).

The introduction of genomic data was made possible using multiple packages from Bioconductor which include GEOquery, hugene10stranscriptcluster.db, oligo, arrayQualityMetrics and the analytics and statistics packages limma, topGO, ReactomePA and clusterProfiler.

Measurement	Normal	Warning	Diabetic
Hemoglobin (%)	<6	6.0 - 6.4	>6.5
Insulin resistance (mg/min)	<4.0	4.1 - 7.4	>7.5

3 Applications

I2EHR contains 6 major components that were deemed most relevant for the integrated analysis of the data: (1) patient tables and observations over time, (2) cohort characterisation by ethnicity, age, sex, disease (3) Cohort observation measurements over time (4) microarray analysis (5) Overlay of the gene expression for an individual gene vs the individuals measurements.

3.1 Individual Patient Query

The individual patient query is accessed by selecting the 'Patient' tab within in the dashboard sidebar. The dashboard is populated with patient ID number "1425bcf6-2853-4c3a-b4c5-64fbe03d43d2" and allows the user to query the information stored for any patient by using their ID code. Using the unique ID number limits the risk of viewing the incorrect patient's information and increases the security of patient data.

Each tab is then populated with information associated with the patients under the headings careplans, conditions, encounters, immunizations, medications, observations, organisations, patients, procedures and providers from the CSV files generated by Synthea. The generated metabolic disorder patients contained no information for the allergies or imaging studies; therefore these were remove for the purpose of presentation. The user can browse the entries within each table passively or use the searchbar to find entries of interest eg. search the dosage of 'simvastatin' prescribed in the medications tab.

```
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])
```

3.2 Plot of clinical observations over time

A plot is located below the patient query datatable containing dates and measurements on the x and y axis respectively.

Enter a figure here...
The dropdown menu is populated by available observation measurements specific to the patient in question. For presentation purposes, the example interface allows selection of only the BMI, insulin resistance and H1bAc levels.

The graph generated will show the values recording throughout time for the individual. When the patient ID is entered above, their name is extracted and populates the plot title. The vertical title on the y-axis will contain the units for the measurement in question; x-axis dates correspond to all dates of patient observations. 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.

3.3 Characterisation of the cohort by co-variates

Cohort level analytics are available within the 'Cohort' section on the dashboard sidebar. The dropdown menu is populated with the covariates age, sex, ethnicity and disease. The default selection is ethnicity. Characterising the individuals by ethnicity is useful for identifying

instances of diseases within an ethnic group. It can also be considered if integrating Genome Wide Association Study results to account for linkage disequilibrium and prevent false discovery of ethnicity-disease associations. Characterising the cohort by age and sex can be used also to understand their association with disease progression. Disease epidemiology can be used to tailor treatment and prevent epidemics.

3.4 Integration of microarray data

Each sample of gene expression data is assigned a patient ID corresponding with a Synthea generated patient.

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.

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 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 samplereadingare considerably different to the rest. 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 assumption s 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 that nature of clinical data collected on EHRs in the FHIR format can be integrated for exploratory purposes and integrated with genomic data to on a single, easily usable platform.

Visualisations of longitudinal clinical data proved useful for identifying changes over time associated with type 2 diabetes 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 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 difficulty 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 core wellness 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 making decisions. 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 that have been noted in the issues section of their GitHub account (ref) and in other reviews. Synthea developers are active on their account and continuously improving the data to stay alike the data available(link).

Another caveat using Synthea occurs when generating control samples. As all samples available are disease patients, each individual has a diagnoses 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 type 2 diabetes 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 for more 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.

Figure 2 Talk about a second use of I2EHR. Talk about how it was implemented through the treatment of type 2 diabetes patients ? might want to know about

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. 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 than 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 single diagnosis of type 2 diabetes. The alternative data source ArrayExpress also showed no results for larger datasets. To address this, values were generated for an expected expression level for a type 2 diabetes patient. This meant that the entire analysis would be a simulation. The complexity of gene expression profiling was a core reason for deciding to use simulated data. As mentioned above, no large type 2 diabetes samples were available online - the largest found consisted of a sample where half also had colon cancer. Given the ambiguity that would be expected with this dataset, it was decided to use simulated values to avoid confusion and continue with a simulation to model the expression that would be characteristic of a diabetes patient.

For app development, Shiny proved especially useful for the development of useful software for the 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 user selected regions of selection to further focus on a certain time in the patients data. 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 upload their raw data file 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.

Acknowledgements

I would like to thank Dr. Pilib Broin for the helpful assistance throughout the project and providing advice 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] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.