OptraHealth
Engineering Digital Healthcare

RWE BASED ANALYTICS FOR DRUG TREATMENT OUTCOMES

OVERVIEW

Payors are increasingly challenging the safety and efficacy of drugs and are demanding real-world evidences as to the patient's conditions and treatment options offered. Based on this growing trend the focus is moving towards a real-world evidence(RWE) approach that aims to use real world live data from disparate data sources. This data is also temporal, thereby allowing for a complete and more precise patient longitudinal view.

The traditional approach to measuring drug efficacy and toxicity during clinical trials and approvals is randomized control trials that randomly matches patients with treatment options wherein patients and provider's choices are not considered and while also allowing for non- adhering patients to opt of clinical trials. Additionally, this approach also introduces a bias by generating a homogeneous patient cohort.

CHALLENGE

The overall success of any drug is vitally dependent up on the patient cohort on which the drug was modeled and later tested and this cohort is a representation of the target population for whom this drug is being developed. This means that defining an accurate patient cohort is one of the most important challenges of drug efficacy and toxicity development and predictions. The challenge we faced was to provide a system that allows for creating a well-defined cohort wherein the cohort definition takes into account all the data that is available for potential members of the cohort.

Typically, only silos of information are used when designing cohorts for e.g. only EMR data. This has limited success but EMR structured data when coupled to other data allows for deeper definitions and thus an enriched view of the cohort.

The challenge we addressed is the development of a system that uses three broad types of data such as temporally active and curated EMR records; secondly, many individuals in a cohort have already undergone molecular testing such as gene expression profiling and or genotyping as part of the primary study, and such data may already exist in the EMR and when it did not we specifically integrated it, lastly, other disparate data available for potential cohort members such as patient registry, social data which may provide additional insights.

Optra HEALTH, *a division of Optra Systems, Inc.*
530 Lakeside Drive, Ste 250, Sunnyvale, CA 94085
Tel: +1-408-524-5300, Fax: +1-408-524-5302, Email: info@optrahealth.com

The second broad challenge was providing the functional needs of each stakeholder in this process. Payors/ organizations require real evidences of patient engagement and for each clinical episode. Clinicians and researchers required analytics level evidences and a platform to drive hypothesis and cohort subjects require information/ data to make informed decisions. It is the ultimate coordination of each of the three stakeholders that lead to using a RWE approach to the fullest and providing the most powerful, insightful and predictive platform for analysis.

Our clients had access several types of data segments described above. Specifically, we developed modules for drug toxicity and efficacy prediction for ovarian cancer using several different types of data. Ovarian cancer is one of the more commonly known aggressive cancers in women and when untreated has a very high mortality rate. Additionally, ovarian cancer presents with vague symptoms such as backache, diarrhea etc. and thereby going undetected until is clinical presentation. At such a time chemotherapeutics are less effective and treatment options have to be planned carefully to minimize the effect of anti-malignants.

SOLUTION

With our many, many years of software expertise and domain experience we developed a platform that is usable across the organization for a certain stakeholder in the process. The following is a list of different assumptions and the strategies that were adopted to address each functional element of the project challenge. The solution is focused on using disparate data from several data sources that provides real word evidences.

DATA SOURCES

In order to establish accurate insights based on real world evidences, an order of magnitude of data sources is required. We identified data sources of different types and which contain different qualitative data that may be efficiently used.

EMR/ EHR

powerful source of data since the data elements are validated and considered of a high quality. However, there are many instances of data in an EMR that are unstructured such as chart & procedure notes. The data elements that we encountered and considered of most importance are:

❖ Patient demographic information	❖ Consultation details
❖ Patient and family disease history	❖ Hospitalization and discharge details
❖ Patient physical report	❖ Patient operative report

We developed modules that efficiently parse EMR data, extracting meaningful information as mentioned above. Advanced parsing and interpretation techniques using machine learning with BioNLP we developed for parsing unstructured data. When molecular level data such as gene expressions was available, it was also used for in the data model.

PATIENT REGISTRIES

We accessed patient registry that had a collection of patient/diseases/therapy-related- information collected through the study method of patients, physicians and laboratory tests. The registry included for health services registries (patients with common procedure/clinical intervention/hospitalization) or disease registries (patients with similar diagnosis). The data elements we used from patient registry are

- ❖ Patient reported data: demographic information, patient reported outcomes (PROs).
- ❖ Clinician reported data: diagnosis, treatment/drug prescribed, laboratory/clinical test suggested, follow-up treatment physician rating of effectiveness.
- ❖ Laboratory: diagnostic/clinical test results.

Patient registry data was also parsed using and also correlated with ontologies and reference databases such as SNOMED-CT. Some of the patients that are represented in the EMR were also represented in the patient registry. This allowed for adding more dimensions to each patient's profile and to include/ exclude certain findings. Additionally, we also developed modules for association wherein a patient was only represented in a registry but not in the EMR. This approach allowed to expand the cohort specially to include patients that have vague symptoms of ovarian cancer.

CLAIMS DATA

Claims data are of a lot of value as they provide insights into medical conditions and diagnosis using ICD and CPT codes. Claims data is another way to confirm and ascertain missing values of data points in an EMR.

We developed modules that are able to accurately identify the medical condition and/ or its severity by correlation with standard databases. This dataset was added to the patient profile and used in cohort definition.

The data elements gathered from hospital claims, provider claims and patient claims are as follows:

❖ Patient claim information. ❖ Patient demographic information. ❖ Consultation details.	❖ Hospitalization details along with cost of treatment. ❖ Diagnoses. ❖ Procedures/drug names with doses and days supplied

SOCIAL DATA

A new emerging trend is the openness with which patients are sharing information about their medical conditions.

We developed connectors and using BioNLP technologies we extracted data from patient oriented sources such as PatientsLikeMe, Twitter, Diabetic Connect. Such social data sites mainly contained unstructured texts, dialogues, sentiments as expressed by the patients or their kin.

The data elements that were gathered from social data are as follows:

- ❖ Unstructured text
- ❖ Dialogues
- ❖ Sentiments

DATA PREPROCESSING

The challenge in acquiring data from the data sources mentioned above was overcome by developing specific connectors. In some instances, data was available as REST API calls but when that was not available we developed compatible connectors that would aggregate the data. The process of data acquisition is simpler than preprocessing it due to prior knowledge of database structure. However, unstructured data from EMR, social data etc. was parsed and meaningful entities identified using BioNLP techniques.

In one implementation of preprocessing, we applied standard vocabularies to harmonize data. For e.g. in one EMR system fever as represented as fever whereas in another case as hyperthermia. Thus using ontologies, we unified both the data to ascertain they are the same. Another such example is data from patient registries and social data where there is consistent lack of control. In such cases even the local culture of writing was understood. This enabled us clean the data, validate and assure its quality. Then this data was made available for downstream analytics.

MACHINE LEARNING & ANALYTICS

Access to curated ovarian cancer data from EMR and by combining data from other disparate data sources we generated a fairly exhaustive database of clinical features. Along with clinical data points we also used molecular data such as gene expression.

Analytics modules were developed for mainly two purposes. First, to generated a detailed view for each cohort and then for each individual in the cohort. Secondly, we developed modules that provided insights into drug toxicity and efficacy for ovarian cancer proving cross linking between clinical, molecular level, side effects and drug toxicity.

PATIENT LONGITUDINAL VIEW

We used all patient data from and generated a 360 deg. patient longitudinal view. This view was deployed as an innovative widget that allows for viewing the patient's entire history as a timeline and allows for easy drill down to any data segment. This view may also be enabled on cohorts.

PREDICTION OF DRUG TOXICITY AND EFFICACY

The first step to model drug toxicity and efficacy is generating precise cohort signatures. Each cohort signature is based on several parameters with each parameter with high/ low weight.

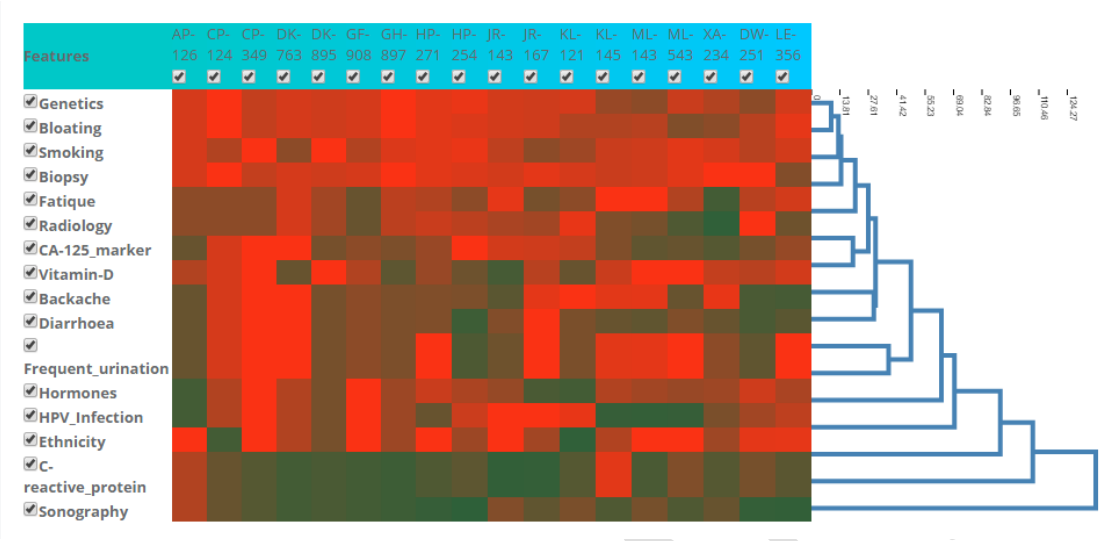

Figure 1: HeatMap of Ovarian CA Clinical Features

Figure 1 is a heatmap generated using clinical features which have high level of confidence. We developed a module that uses a supervised machine learning approach where each clinical feature served as one measurable label. The algorithm then analyzed each patient sample for each clinical features and by applying clustering, we grouped patients with similar features in clusters. The color on the HeatMap are intensity representation of the measured features.

The module allows for the algorithm to be tweaked to identify features with missing values to exclude them from the analysis. So the result of this technique is grouping patients with similar clinical profiles into clusters such that each clusters has a unique signature.

For downstream analysis such HeatMap and cluster data is saved in a database and this data serves as the training set for the machine learning algorithms. This allows us to test any new random patient profile, and if the profiles' signature matches the signature of any one sample then we predict that the random sample also has similar clinical outcomes and cellular biochemical properties.

Such a system, wherein a large reference dataset is used to generate precise cohort signatures forms the basis of a prediction system for early identification of cases (with vague symptoms) and also for accurate disease staging.

Then we extended our current system to generate cluster HeatMaps that included gene expression profiling. Using drug- gene interaction databases we generated a concept map as shown in Figure 2

which show the interactions of known drugs with genes. Doxobubicin, Cisplatin, Paclitaxel are the commonly used anti-malignant in treatment of ovarian cancer and Figure 2 specifically outlines the interactions of each of these drugs with genes.

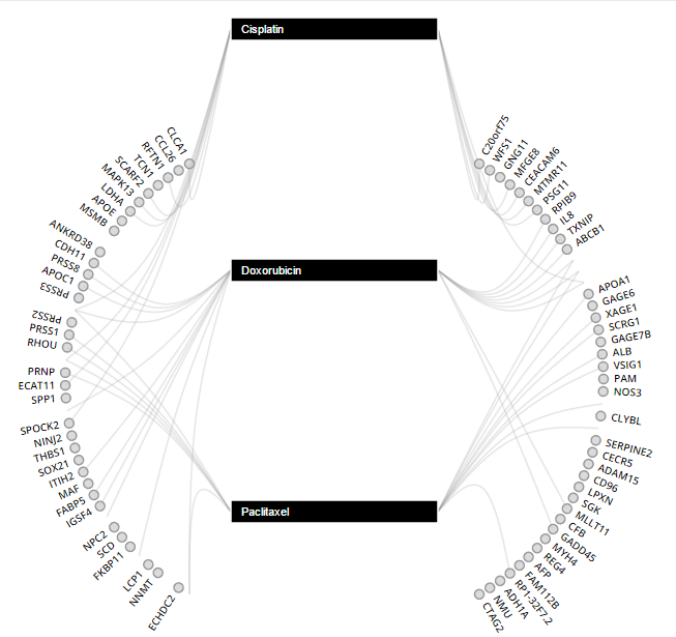

Figure 2: Drug- Gene Interactions Concept Map

Since the ovarian cancer dataset also contained genetic data, we identified up and down regulated genes and mapped them to drugs using our software modules. By combining clinical data heatmap with gene expression heatmap we have a system that provides us molecular insights.

We tested the model using random patient profiles and executed our methods over it. The algorithms classify the test sample to its closest cohort signature. Our software tools allowed us to interrogate gene expression profile of any cohort providing drug interactions information. Since our random test sample is classified to any one cohort signature, the interactions of the sample with drugs would be similar to the interactions of the cohort with specific drugs.

BENEFITS

We have developed a platform and its associated module that overcome the typical problem encountered by organizations in the late and mid drug development. Accurate cohort definitions, informed inclusion and

exclusion criteria, modeling of drug toxicity, efficacy and effects with co-morbidity are most commonly encountered.

While some of the problems may be addressed by connecting to structured databases such as EMRs and some others are overcome by accessing clinical trials data. Such trial is at times not directly application and therefore introduces randomness. To overcome such issues and provide deeper insights into real world connection to actual patient data from structure and unstructured datasources and of disparate natures makes clinical trials more meaningful and insightful.

SCALABILITY **Scale from a single server to many servers automatically to balance load and reduce analysis time.**	EXTENSIBILITY **Use any code that you have developed and call that with our agnostic modules.**
ON- DEMAND **Enable/ disable modules for your users on the fly.**	DOMAIN WRAPPERS **We provide ready end-to-end modules for e.g. for tumor- normal analysis, drug interactions, population studies etc.**
COMPLEX VISUALIZATIONS **View complex datasets using the latest technological tools and with recommendations of widgets for data.**	KNOWLEDGE AUTOMATION **Turn data into knowledge by leveraging our BioNLP powered cloud data repository.**
INTEGRATION **All modules are available as REST APIs allowing you to integrate with your solutions.**	