synergus RWE

evidence for decisions

Data Science in Oncology
Internship summary

This report has been compiled by

Laila Mehkri, MD, MPH, Health Economics, Project

Contents

Table of Contents

1 Introduction

1.1 The Evolving Availability of Data to Support New Solutions

1.1.1 Big Data

The concept of “big data” appeared in the early 2000s. There is no standard definition for big data, but generally refers to large amounts of information (massive and complex data sets) too large for human analysis or to adequately manage and process with traditional software (Deng and Xing, 2019). Initially big data was defined as the three Vs: for volume, velocity and variety of information. A 4th V was added later, referring to veracity, which means reliability of the accumulated data (Barbosa, 2017).

The pace of data generation has increased, storing huge volumes of data, to effectively manage, analyze the information collected to support new solutions is part of the challenge. The ability to gain insight from within these vast depots of data has become the new challenge in Oncology to support new solutions in the field (Fessele, 2018)

The heterogeneity in oncology disease manifestations, various treatment options available, patients’ preferences and deciding which treatment is optimal for an individual patient are some of the most significant challenges. To make reliable decisions, large amounts of data is required (Osong et al., 2019)

According to IBM, approximately 2.5 quintillion bytes of digital data are generated from humans’ daily activities (e.g. from groceries, demographic and administrative medical records) (Petrov, n.d.). The acquisition or extraction of these extensive and voluminous data is colloquially known as big data.

Big data research is data driven: methods may first be applied to the data itself to identify potential causal relationships. This can result in a list of associations with varying degrees of correlation that can then be further evaluated. In this way, big data analysis may start the research process before identifying the important questions. Big data analytic techniques include data mining and machine learning, (Berger and Doban, 2014) which offer potential alternative approaches to leveraging large data resources (Trifiletti and Showalter, 2015).

1.1.2 Genomics

Genomic involves many sciences including genetics, molecular biology, biochemistry, statistics, and computer sciences. Gregor Mendel put forward the foundation of modern genetics with the discovery of the basic principles of heredity. This has allowed huge amounts of genomic data to be collected around the world by different organizations. AstraZeneca, which is one of the world’s largest pharmaceutical companies, compiled genome sequences and genomic health records from over 2 million people. The company and its collaborators wanted to uncover rare genetic sequences associated with diseases and with responses to treatment (Srkalovic, 2019).

The need for sequencing along with the study of various genetic diseases and drug development for the diseases has become an important part of the oncology practice. With the advancements in technologies, genomic data can be easily collected at a faster pace and at a lower cost. The importance of individual genome sequencing is that it supports personalized medicine.6 This era

of personalized cancer medicine based on the molecular characteristics of an individual patients tumor has great potential in the therapy of any type of cancer (Srkalovic, 2019).

Development of targeted therapies for some cancers has been done by the discovery of single gene mutations (genetics) and some cancers rely on expression profiles composed of many genes (genomics). For examples in breast cancer, advances using genomic data have been driven by expression profiles more than single gene mutations. Although breast cancer is complex and heterogeneous, the field shifted substantially with the publication of microarray-based gene expression profiles demonstrating the heterogeneous nature of breast cancer at the molecular level and the sub classification of breast cancer into different molecular subtypes (Trivedi et al., 2019).

Previously only the clinical-pathological factors (patient age, tumor size, histopathology features, lymph node involvement) were used to estimate probability of breast cancer recurrence in patients; but now with the evolving availability of data there has been a substantial interest in developing molecular assays that more accurately predict clinical outcome in a selection of patients who will most benefit from more aggressive therapies and on the other hand avoiding overtreatment in those patients with comparatively low risk of distant recurrence of breast cancer. This means that the selection of a combination of genes that provide information about a tumor's metastatic potential will most accurately predict distant recurrence in that patient (Trivedi et al., 2019).

2 Brief Summary of Current Evidence

2.1 Data science for Drug discovery and development and finding match target compounds for molecular profiles

With the vast variety of data sets being generated in the clinic the drug discovery and development goes through many difficulties and benefits associated with it. Drug discoveries and development not only depends on the clinical data but also depends on other data sources and disciplines such as chemistry, pharmacology and bioassays, mechanistic mapping of disease and structural biology. The big data which is derived clinically and informs about the mechanisms of recurrence and resistance to therapy is already in use today for driving new drug development aimed to address unmet clinical needs (Workman et al., 2019).

Drug discovery and development is evolving rapidly in the world today, with novel drugs supported by the use Big Data building knowledge and gaining information from previous successes and failures (Sellwood et al., 2018). The prioritization and validation of the selected targets is the first step for drug development and discovery. It is done by using both biological and chemical big data (Workman and Al-Lazikani, 2013).

The largest sources of clinically- derived information on genomic alterations in cancer, disease drivers and pathogenic pathways are The Cancer Genome Atlas (TCGA) ("The future of cancer genomics," 2015) and The International Cancer Genome Consortium (ICGC) (Zhang et al., 2019). The use of chemical and pharmacological resources has become very important for drug discovery and development because of the extensive amounts of prior knowledge in them which is important in informing the screening drug development campaigns and chemical library design (Mullard, 2017). ChEMBL is the gold standard database of experimentally determined bioactivities and small molecules. PubChem is another database, which is much wider and is community-driven. It contains chemical data on millions of compounds with associated bioassay results, which are used for drug development

processes. The data sources on absorption, distribution, metabolism, excretion and toxicity (ADMET) information are limited but the major chemical/pharmacological databases like ChEMBL and PubChem contain some small data on ADMET (Gaulton et al., 2017; Wang et al., 2017). These available public resources are invaluable for drug development campaign, but they are small and very diverse which makes it a challenge for drug discovery and development companies to apply them optimally and effectively in the context of their projects.

Integration of information is also very important for drug development decisions but it is also a challenge as very few resources are available. One such resource is canSAR, which is the world's largest public resource for cancer drug discovery and development. It integrates data from the clinic, genomics, medicinal chemistry, pharmacology and 3D protein structure and refines all this into single integrated reports to help the drug discoverers. This resource can be used to help in evidence-based assessment and prioritization of targets for drug discoveries and developments (Coker et al., 2019).

It is also very important to understand the machine learning and artificial intelligence approaches in drug discoveries and developments. Artificial intelligence or AI refers to a computer algorithm that can 'learn' patterns from input data and apply this learning to make novel predictions from new data. Machine learning and Deep learning also falls into this overall definition (LeCun et al., 2015). The information that the researchers like to access and refine is hidden within vast, sparse data. For example, such data include all chemical compounds, atomic interactions and synthetic reaction pathways in medicinal chemistry. It is in these huge-scale Big Data areas that complex deep learning algorithms come into their own. In contrast to machine learning, deep learning can discover key patterns hidden in a multidimensional data space through the layered abstraction of diverse data (LeCun et al., 2015).

The canSAR resource for cancer drug discovery and development asseses target manageability based on protein 3D structure, chemical properties of bioactive compounds and the structure of cellular networks by utilizing machine learning approaches (Mitsopoulos et al., 2015; Patel et al., 2013). The mapping of drug development opportunities on the disease-causing molecular networks, which are identified from systematic profiling of cancer patients, has now been possible by the application of such integrated machine learning capabilities. Not only this, but these approaches help in future drug discovery and development campaigns (Wedge et al., 2018).

The complexity and immensity of medicinal chemistry space naturally lends itself to the use of data science technologies for the creation of novel drugs. Medicinal chemists apply artificial intelligence methods to the design and prioritization of compounds for drug development. The application of artificial intelligence in this area with the data availability is increasing and the field is now maturing more as more and more academic and industrial laboratories are now utilizing this technology for drug designing and development (Sellwood et al., 2018). Artificial intelligence when used with the experience and creativity of the medicinal chemists, can reduce the tests cycles for drug designs and new drugs are discovered more rapidly with the use of fewer overall resources (Workman et al., 2019).

In the field of bio therapeutics, deep learning algorithms are being used in an effort to better inform immunotherapeutic development and it is also helping in designing of novel antibodies which shows how bio therapeutic design is also benefiting from Big Data and Artificial intelligence (Bulik-Sullivan et al., 2018; Nimrod et al., 2018).

Cancer-therapeutics startups have grown from $2bn in 2013 to $4.5bn in 2017 according to equity investment. Not only big pharmaceutical companies but also technology companies like Google and IBM have recently invested millions into startups, which are developing immune-oncology and targeted therapy treatments. All cancer drugs that are developed take up fully 31% of all drug development programs and have overall success rate of 5.1% of FDA approval. The whole drug development process takes almost 10 to 15 years with $2.6 billion only to bring a cancer drug to the

market. All this shows that a big dose of digital transformation and use of data science is required as many of the analytical processes involved in the process can be made more effective and efficient with artificial intelligence and machine learning. This has made it easier to save not only years of work but also hundreds of millions in investments (Alamanou, 2019).

Atomise is drug molecular discovery focused company advancing AI technology by using deep learning to analyze molecules through simulation that helps in saving time taken by the researchers to synthesize and test compounds for drug developments and discoveries. The company has over 50 different molecular discovery programs and helps to discover molecules that help prevent diseases while also studying their toxicity and their reaction with the human body (Alamanou, 2019).

GlaxoSmithKline (GSK) formed a collaboration with an Artificial intelligence driven company Insilico Medicine in order to explore how Insilico's artificial intelligence technology can help in the faster cancer drug identification of novel biological targets and pathways of interest to GSK. Numerate is another data driven drug design company applying Artificial intelligence to drug discovery and leading discovery programs for identifying clinical candidates in Takeda's core therapeutic areas of oncology. Another partnership between Crystal Genomics and Artificial intelligence driven company Standigm is working towards discovering and developing novel drugs to treat cancer and liver related diseases (Alamanou, 2019).

2.2 Data science use in clinical trial recruitment

Deep 6 AI is an Artificial intelligence driven company that uses Artificial intelligence to extract valuable information and medical records to speed up the process of finding and recruiting patients for clinical trials within minutes. In one comparison it was shown how Deep 6 artificial intelligence software found and validated 58 desirable matches in less than 10 minutes in comparison with only 23 desirable patients found in six months for a biomarker for non small cell lung cancer trial by principal investigator which used traditional recruitment methods. Deep 6 AI, uses Natural Language Processing (NLP) in order to read doctors' notes, lab reports, diagnoses of the patients, recommendations, and to detect lifestyle data which is hard to find, such as smoking and activity history (Alamanou, 2019).

NVIDIA is an American technology company that is using Artificial intelligence in creating a discovery platform for cancer called CANDLE. It's aim is to uncover the genetic DNA and RNA of common cancers, also help in predicting how patients will respond to treatments, how each individual patient's cancer evolves and transforms, it will also help in building a database of cancer disease metastasis and recurrence, and aims to get the new therapies to the market faster (Alamanou, 2019).

The clinical trial companies that are using Artificial intelligence are facing three major challenges. One is the capability of accessing patient data, second is accessing good quality data and third challenge is to create an industry-wide data standards. The data standards need to have patient data from a vast range of sources including mobile devices, wearables, data from healthy populations and not just be restricted to only cancer patients. All this can be helpful in identifying and understanding spatial patterns, behaviors and life style (Alamanou, 2019).

2.3 Impact of Data Science On Precision Medicine

Precision or personalized medicine is a medical approach in which patients are classified or arranged based on their disease subtype, risk, prognosis, or treatment response using specialized diagnostic

tests and tools (Sobradillo et al., 2011). The main idea of this approach is to base medical decisions on individual patient characteristics that also include the biomarkers (e.g., genomic alterations, molecular markers, disease severity scores, lifestyle characteristics, etc) rather than on averages over a whole population. Personalized medicine helps with better medication effectiveness, since the treatments are tailored to according to the patient characteristics, e.g., their genetic profile. It helps in the reduction of adverse events and risks through the avoidance of therapies and drugs that shows no clear positive effect on the disease but at the same time those therapies exhibit negative side effects that maybe partially unavoidable (Vogenberg et al., 2010). The impact of data science on precision medicine has also helped in lowering the healthcare costs by optimizing and effectively using the therapies (Fröhlich et al., 2018). It has also helped in early disease diagnosis and prevention by using molecular and non-molecular biomarkers (Jensen et al., 2012). Use of precision medicine has also improved disease management with the help of patient wearable sensors and mobile health applications (Lee and Yoon, 2017) and has by providing with smarter designs of clinical trials by selecting the likely responders at baseline.

Modern approaches from data science, specifically machine learning, have a considerable impact on personalized medicine. There are some identified challenges that cause hurdles for realizing machine learning-based solutions more broadly in clinical practice and more so when focused on precision medicine. There is existing potential of data science for personalized medicine and many directions for its future development.

Personalized medicine is tightly connected with genomics. Genomics and other biological high throughput data for example transcriptomics, epigenomics, proteomics, metabolomics are not the only source of data used in the personalized medicine field. Other relevant data includes bio-images (MRT and CT scans), electronic medical records (EMRs) (Fröhlich et al., 2018), health claims data from insurance companies (Jensen et al., 2012), and data from patient wearable sensors and mobile health applications (Lee and Yoon, 2017).

It is important to understand that clinical drug responses are highly multi-faceted and dependent on combination of patient intrinsic (e.g., genomic, age, sex, co-medications, liver function) and extrinsic (e.g., alcohol consumption, diet, sunlight exposure) factors (Fröhlich et al., 2018). Multivariate classification algorithms use techniques from the area of Artificial Intelligence (including machine learning) that play an important role in precision medicine. MammaPrint™, is a highly cited example, it is a prognostic test for breast cancer based on a 70-gene signature (van ’t Veer et al., 2002). MammaPrint™ was approved by the FDA in 2007 and produces a score from the weighted average of 70 measured genes, which helps in the prediction for the development of distant metastases in patients. By the addition of MammaPrint™ signature in clinical practice for precision medicine, its utility can be compared to standard clinicopathological criteria as it helps in selecting patients for adjuvant chemotherapy (Cardoso et al., 2016).

2.4 Supporting the complex decision making of the physicians/oncologists

In oncology the use of data science has become important part of evidence-based decision-making, where integration and analysis of massive datasets aims to improve the quality, efficiency, cost and outcome of important business decisions in oncology (Taglang and Jackson, 2016). Supporting the complex decisions by using data science in oncology for the physicians can mean differentiating between toxic and efficacious treatment choices for the patients and improving patient quality of life and longevity. The use of data science is garnering increasing awareness in oncology where the concept is empowering a paradigm shift away from generic standards of care, towards personalized

approaches to prevention, early detection, early intervention and optimized disease management for cancer patients.

In oncology significant amount of investments have generated a diversity of quantitative (e.g. Next Generation Sequence (NGS) data, sensor data, imaging information, laboratory tests) and qualitative (e.g. demographics, doctors letters) data-types for individual patients. Integration of these datasets consisting of imaging, molecular, genetic/genomic, cellular, organismal, environmental, family history and lifestyle data helps in truly personalizing the medicine. Synthesizing all of this into an actionable recommendation requires immense data processing capacity. Such datasets enable the concept of "the digital patient", a revolution that provides fertile ground for new innovations in knowledge discovery and data-driven clinical decision-making.

IBM Research announced in 2007 that the computer IBM Watson was taking on medical science. Watson applies its DeepQA open-domain question answering technology to provide an evidence-based clinical decision support system. Watson uses natural language processing and machine learning to analyze unstructured information, overcoming the challenges with the structured data of traditional expert systems. It generates a list of possible questions and uses abductive reasoning to generate hypotheses and produce possible answers from the available information (Trivedi et al., 2019).

Flatiron, a USA based healthcare technology and services company focused on accelerating oncology research and improving patient care. They have developed an integrated cloud-based platform called OncoEMR which optimizes clinical decisions and reduces the costs. It provides all clinical and patient information needed to make quality care decisions and diagnoses, and all patient information can be made available in one place. It allows all the clinical summary, documents and physician notes are integrated to streamline physician workflows. Its ability for inbound and outbound faxing helps make the communication with referring physicians faster and easier, providing patient updates and documentation in minutes. There is a patient portal for the patients integrated with OncoEMR, that helps the patients to view their charts and lab results, and communicate with the practice securely online, results in reduced calls to the front desk. Its mission is to help reduce the costs as well due to elimination of medical record supplies and hardware costs, and reduction of contract staff. This cloud-based platform has given faster and easier access to clinical and patient data resulting in streamlined workflows for physicians and staff and improved the patient experience ("Flatiron Health," n.d.).

3 Limitations and Challenges with the use of Data science in Oncology

One of the challenges with data science use in oncology practice is that it requires experience in advanced analysis techniques and the new vocabularies that come with them. This has made it impractical for physicians to gain expertise in these techniques and they require assistance of public health and computer science experts who know how to work around computational hardware which is capable of such demanding processing and mathematical modeling. This requires the researchers and the physicians to have the knowledge of these techniques in order to interpret the resulting publications. In order to realize the potential of a learning health system, a large number of analysts and researchers require training in big data analytics and health care information technology (Trifiletti and Showalter, 2015).

Ownership of the data serves as another limitation with data science use in oncology. It has been estimated that 91% of patients permit their personal health records and information to be shared for medical research purposes, however the collection and management of the data is not performed by

patients so there is a need of data management companies in big data research.
Here however, privacy plays an important consideration that work as a limitation in the development of big data research, since data breaches could harm patients and prevent health system participations in big data exchanges due to fear of data security compromises. The role of informed consent documents and institutional review boards (IRBs) to protect patient rights in big data research is unclear but will likely be defined in the coming years (Gupta, 2013)

There can be concerns regarding the validity of the predictive algorithms and methods, which may come across as another challenge as the application of big data predictive analytic methods help in making health care decisions. It makes it important to be focused on validating predictive models. Clinical application of big data analytics requires trust in these methodologies on the part of clinicians, with commensurate efforts to evaluate the performance of decision support tools and information provided by big data in health care. As a result, predictive analytic models using big data to guide health care choices must be developed and executed in a transparent and replicable way with validation in real-world conditions (Cohen et al., 2014).

Another challenge is the limited prediction performance for clinical practice of the machine learning models in data science, this is due to the descriptive power of the employed data with respect to the clinical endpoint of interest. Many of the machines learning models that are developed for personalized medicine do not have a predictive power for clinical outcomes close to the high level and the unrealistic expectations of clinicians. There are certain reasons for this which act as limitations. The relationships of patient-specific characteristics to clinically relevant endpoints are very complex since they are not well described by one data alone and often vary over time. It is also a challenge to discriminate relevant patient specific features from irrelevant features specifically when it comes to high throughput genomics data. Another challenge is that the clinical outcomes may be partially influenced by factors that are hard to capture (e.g., social and environmental influences). It is also a challenge to obtain a large patient cohort with well-defined phenotypes for testing models due to the cost and time limitations (Fröhlich et al., 2018).

Many of the genomics data is very noisy due to technical measurement errors and due to being highly informative with too much biological variations in the data. There have been however no advances in methods to discriminate these kinds of noises. Machine learning models are sensitive to selection biases and this impacts the prediction performance aswell as there is no careful choice of patient samples. The models may under or over represent specific patient subgroups in the testing cohort and there are also some ethical considerations which have not been explored yet. Example could this be a racist prediction model, due to under or over representation of certain ethnicities (Caliskan et al., 2017). Careful designing of the training set is important to ensure that it represents the patient population in the intended application phase of the model in clinical practice (Fröhlich et al., 2018).

Another limitation regarding AI or machine learning techniques is the difficulties in interpretation, that it doesn't provide deeper or causal understanding of an observed phenomenon, however it can detect complex patterns in large data sets and provide accurate predictions as well. They can only capture statistical dependencies, correlations from the data sets but the use of data science and AI do not replace the classical hypothesis driven researches. The lack of clear causal interpretation of machine learning models hinders the acceptance of data science-based solution by the physicians even if an acceptable prediction performance is achieved (Fröhlich et al., 2018).

Another limitation with Data science use in oncology is the Insufficient validation for clinical practice. There is an emphasis given on the requirement of validation when establishing any algorithm for patient stratification in clinical practice. The quality of how good a machine learning model to the training data is or to the training error is not indicative of its later performance on unseen data and is very overly optimistic. Validation for clinical practice has several steps, as follows: Internal validation is based on the initial discovery cohort and this can be done by setting data parts aside via cross-

validation or as an independent test set. Cross-validation is a method by which a fraction or percentage of the original data is left out for model testing and the remaining of the fraction is used for model training. Cross-validation strategy averages the prediction performance over different test sets and reduces the variations in the estimation of the test set performances. This can be particularly relevant in a not very large discovery cohort. Another step for validating for clinical practice is the external validation strategy, which is based on an independent cohort. This addresses the potential selection bias, which is a necessity during the compilation of the discovery cohort. The last step would be to the validation of clinical practice in a prospective clinical trial to demonstrate and show the benefits of it compared to standard care. The limitation in this is that this entire process is costly and time consuming and the number of validated models for clinical practice are also limited (Cardoso et al., 2016).

Existing challenges and limitations with data science and AI can be summarized as the insufficiency in prediction performance, the challenges with model interpretation and the validation of the stratification algorithms into clinical practice. There are however high expectations portrayed in the mainstream media regarding the use of Data science and AI not only in the field of oncology but also healthcare in general. However, there are only few machine-learning based solutions available, which are used by the companies that impact the clinical practice. These challenges are however pointed out in more detail in later sections and the possible ways of how these are addressed currently and will be addressed in the future.

4 Main Section

A great number of companies and other stakeholders are now moving to the use of real-world data (RWD) to amplify the use of RCTs in oncology as well as in other therapeutic areas. A commonly accepted view in Europe of RWD is that it constitutes “longitudinal patient level data captured in the routine management of patients that which can be repurposed to study the impact of healthcare interventions”. These includes Electronic health record (EHR) data on patient symptoms, referrals, prescriptions and treatment outcomes (including patient-reported outcomes [PROs]); Electronic medical records (EMR), Claims data on service usage, insurance and other administrative hospital data; Genomics databases and associated biomarker data; Drug databases also known as Pharmaceutical databases. Social media data, for example from patient forums; Mobile applications data, wearables and sensors data; (Montouchet et al., n.d.). These also include Cancer registry data, clinical trial data, epidemiological data, scientific literature, patient reported feedbacks and chemotherapy regimen data.

Companies that are using Electronic medical records includes Berg, Concerto health, Cota healthcare, Flatiron, IBM Watson healthcare, RTI health solutions, Sharp analytics. Companies involved in the use of Electronic health records include Aetion, Clinerion and Oncology analytics. Aetion and RTI health solutions are also involved in the use of Claims data along with Flatiron and in the use of cancer registries along with Lancor scientific.

Flatiron uses clinical trial data along with IBM Watson healthcare, RTI health solutions, Sophia genetics, Concerto health and Cota healthcare. Cota healthcare along with Oncology analytics and Pubgene use cancer regimen data also.

A number of companies are involved in the use of Genomic databases such as Berg, Cota healthcare, Exploris, IBM Watson healthcare, Oncology ventures, Pubgene, Regeneron and Sophia genetics. Regeneron is also involved in the use of scientific literature.

4.1 Electronic medical records (EMR)

Electronic medical records (EMRs) are digital versions of paper charts in the hospitals. An EMR of a patient contains all the medical and treatment history in one place. EMRs have advantages over paper records in the healthcare facilities as they allow the healthcare professionals to track data over time, help in identifying patients who are due for preventive screenings or checkups, check patients parameters like blood pressure and vaccinations, helps in the monitoring and improvement of the overall quality of care with the healthcare practice (Montouchet et al., n.d.).

In a patient's EMR, there are laboratory values, diagnostic reports, radiologic image sequences (every pixel), and clinical notes full of dictation errors and misspellings. The information provided in big data sets can be grouped into structured data (numerical laboratory values for example) and unstructured data (a physician's clinical impression text). Healthcare big data is viewed as a big clinical database like any other clinical databases (Trifiletti and Showalter, 2015).

4.2 Electronic Health records (EHR)

Electronic health records (EHRs) are the same as EMRs but they focus on the total health of the patient and not only the standard clinical data collected in the physician's office. EHRs reach out more than the health organization that originally collects and compiles the information. They are made to share the information with other health care providers, such as laboratories and specialists, and not just stay restricted to the organization that collects and compiles them, this way they contain all the information from the clinicians involved in patient's care. According to the National Alliance for Health Information Technology, EHR data "can be created, managed, and consulted by authorized clinicians and staff across more than one healthcare organization." ("EMR vs EHR – What is the Difference? | Health IT Buzz," n.d.).

According to HIMSS Analytics, "The EHR represents the ability to easily share medical information among stakeholders and to have a patient's information follow him or her through the various modalities of care engaged by that individual." EHRs are designed to be accessed by the patients themselves as well as by all people involved in the patients care ("EMR vs EHR – What is the Difference? | Health IT Buzz," n.d.).

In the year 2009, the "Health Information Technology for Economic and Clinical Health (HITECH) Act" was enacted to support the use of electronic health records - meaningful use [EHR-MU], an effort led by Centers for Medicare & Medicaid Services (CMS) and the Office of the National Coordinator for Health IT (ONC) (Henricks, 2011). By the increasing adoption of electronic health records (EHRs) in drug development and clinical trial recruitment, has promoted its meaningful use intended to increase data capture and sharing, improving the clinical workflow, and eventually demonstration of improved patient outcomes (Henricks, 2011). This meaningful use implies the use of certified EHR technology in a manner that provides for the electronic exchange of health information to improve the quality of care by supporting clinical trials and big data driven drug developments and discoveries (Chen and Snyder, 2013).

4.3 Claims data description

Claims databases have claim history of all outpatient prescription pharmacy services provided and covered by the health services and plans. Drug claims data include mane of the drug, dosage form of the drug, strength of the drug, filling date, days of supply of the drug, financial information of the patient, and de-identified patient and prescriber codes, that allow longitudinal tracking of medication refill patterns and changes in medications of the patient.
Collection of medical claims is done from all the available health care sites (inpatient hospital, outpatient hospital, ER, physician's office, surgery center, etc.) for all types of services being provided, including specialty, preventive and office-based treatments. Claims for ambulatory services submitted by individual providers, e.g., physicians, use the CMS-1500 formats. Claims for facility services such as hospitals, use the UB-04, or CMS-1450 formats.

Medical claims include: multiple diagnosis codes recorded with the ICD-9-CM diagnosis codes; procedures recorded with ICD-9-CM procedure codes, CPT or HCPCS codes; site of service codes; provider specialty codes; revenue codes (for facilities); paid amounts; and other information. Typically, facility claims do not include medications dispensed in hospital (Birnbaum et al., 1999).

4.4 Genomic Databases

Genome databases are organized collection of all the information from the mapping and production of genome (sequence) or genome product (transcript, protein) information. They contain a variety of biological information. The making of genome databases involve acquiring information that researchers have generated and then organizing it into a database in order to make biological inferences. Genome databases vary widely and reflect the communities that they serve (Hide, 2005).

The Genome Database (GDB) is for the human genome, originally developed at Johns Hopkins University in Baltimore, Maryland. Since 1990 it has been one official central repository for genomic mapping data in Homo sapiens but now it also contains sequence information, which was not its original focus. Genome databases have tended to contain significantly greater proportions of genome sequence as the field is moving rapidly from basic genetics and mapping into sequencing era. NCBI and Ensembl, funded by the Wellcome Trust, are the two largest genomic database systems and provide the most comprehensive resources. Both the systems are ‘sequence-centric’. These systems have the capacity and the ability to support several species of genomes and have developed tools and systems for analysis of human, mouse and several other organisms (Hide, 2005).

4.5 Drug Databases

Drug databases are also known as tertiary online drug information sources. They provide broad set of information on pharmacology, efficacy, safety and other topics related to drugs therapeutic use and properties. Due to their online availability and user-friendly platforms, where the content is updated faster, the drug databases are very easy and fast to access. They are able to easily provide valid evidence from the latest research at the point of care (Silva et al., 2013).

For example the FDA drug database which includes most of the drugs approved in the U.S. since 1939. To search this database, you need to go on the FDA drug databases website and type in the name of the drug or any active ingredient of a drug. The FDA drug database can also be used to search drugs that are currently going through clinical trials and/or the approval process. The drugs that have not

been submitted to the FDA but not yet approved can be found in this database ("Pharmacy Drug Databases & Web Resources," n.d.).

4.6 Social Media Data

Social media is a product of information technology (IT) and is also facilitated by it. The use of all the information from big data has been very useful in the areas of health and medicine for gaining deeper insights on the risk behaviors, disease outbreaks and how these can be managed by monitoring the social media and big data activities (Sagiroglu and Sinanc, 2013; Young, 2014). Social media allows users to connect, communicate and interact with each other through which big data is generated and this makes it important to be monitored and controlled as there is high volume of use engagements (Correa et al., 2010; Mgudlwa and Iyamu, 2018; Young, 2014)

Social media consists of internet-based channels that allow its users to interact and share information with eachother without any geographical boundaries or limitations (Carr and Hayes, 2015). Big data has a lot of importance to businesses and it is also very interesting and challenging to see how this massive volume of information is being stored, shared and managed (Kwon et al., 2014; Yin et al., 2012)

Social media produces big data that has a lot of importance for the healthcare sector companies and organizations, but this may create challenges when considering privacy in the healthcare environment (Katal et al., 2013).

4.7 Mobile Applications Database

Medical applications used in mobile devices like smartphones and tablet computers now have advanced and providing specialized tools and resources to the healthcare professionals, patients, and the public (J Seabrook et al., 2014). There has been a rapid increase in the use of mobile applications in clinical practice. Health care professionals use medical devices and applications for administration, health record maintenance and access, communications and consulting, reference and information gathering, and medical education (Ventola, 2014). These faster processing devices, efficient open source operating systems which are able to perform complex functions have made their way into medical professional and personal use and made clinical resources available easily to the healthcare professionals (Ozdalga et al., 2012; Payne et al., 2012).

Medical applications are available for a lot of different purposes, like electronically prescribing medicines to the patients, diagnosis and treatment, practice management of the healthcare professionals, coding and billing, e-learning for the professionals, drug reference guides, clinical guidelines, decision support aids, literature search portals as well as applications that simulate surgical procedures and conduct simple hearing and vision examinations (Mickan et al., 2013; Mosa et al., 2012; O'Neill et al., 2013; Payne et al., 2012). In 2013, the purposes were further divided into categories that also included Electronic medical records, patient monitoring & education, imaging and personal care ("Apple helps MDs cut thru medical apps clutter," 2011)

4.8 Wearables and Sensors Data

Information and communication technology (ICT) has now been advanced with tele-consultation and tele-surgery applications to support the wellness and independence of patients. Along with this sensors are also embedded in patient's environment that can provide continuous monitoring of the patient's activities at home. These however are still not able to sense physiological information of the patients (Yamada and Lopez, 2012).

Wireless communication systems in wifi, Bluetooth, near field communication have been adopted in healthcare and mobile devices. Since 2007 there has been an advancement of information processing platform for the patients that is to be used for the continuous monitoring of their everyday healthcare data so people can know about their health by themselves. These wearable physiological sensors have wireless networking capabilities that help them arrange the collected data. The database technology for this sort of information provides high security and privacy and enables the collection and accumulation of the massive amounts of data to be shared efficiently. This approach also enables the individuals to give feedback for better and efficient services (Yamada and Lopez, 2012).

4.9 Cancer Registry Data

Census of cancer cases is provided by cancer registry data and it is used to monitor cancer incidence at the local, state and national levels and to evaluate and investigate the effectiveness of public health interventions and the cancer treatment patterns. Cancer registry data works as a stable foundation by the public healthcare sectors to reduce the cancer incidences, mortality and survival (White et al., 2017). It is designed for the collection, management and proper storage of data on people with cancer and plays an important role in cancer surveillance to improve efforts to reduce cancer burden. It may also serve for cancer research and for cancer prevention and control interventions.

There are three types of cancer registries that are population based registries, hospital based registries and special cancer registries. Population based registries have records of population in a geographical area and to use the data for epidemiology and for public health purposes. They determine cancer in various populations, monitor different cancer trends over time in a population, is used in different health services and epidemiological researches. This helps the public health providers to prioritize health resource allocations effectively and efficiently. On the other hand the hospital based registries maintain the data of all the patients diagnosed and treated for cancer at a specific healthcare facility. This helps the healthcare facility improve patient care, do clinical research and use the information for professional education. Special cancer registries have data on a particular type of cancer and these provide information to research on specific cancers or provide support to those who suffer from it (“What is a Cancer Registry?,” n.d.)

4.10 Clinical trial Data

Clinical trial data serves as an important resource for medical and health research. This data is collected as part of a formal clinical trial program. There are different clinical trial databases that have different goals. During the course of a the entire clinical trial process, different types of data are collected, monitored and then transformed into data sets that are more analyzable for different research purposes, or to generate various publications and reports for different organizations (Drazen, 2002). For example Clinicaltrials.gov have information on public and privately supported clinical studies from around the world or CenterWatch which is a portal for actively recruiting pharmaceutical industry sponsored clinical trials (Rich, n.d.).

4.11 Epidemiological data

Epidemiological data refers to various forms of nonexperimental observations that include population exposure levels and health effect values and are observed from the samples. There are obvious errors in estimating human risks from animal experiment data and the limitations clinical case reports have when discussing about the entire population perspective and this is where the epidemiological data helps as an evidence for a risk assessment (Ni et al., 2017). The purpose of the data is to inform decisions about the control of health problems. The entire population perspective in the data allows us to be interested in looking at the overall average affects in the groups that are being investigated. This data is essential for further research but difficult to guide clinical decisions. As the clinicians require

information about the specific risks and benefits faced by the individual patients and not based on average effects (Hannaford and Owen-Smith, 1998).

4.12 Scientific literature

A scientific database is a computerized, organized collection of related data, which can be accessed for scientific inquiry and long-term stewardship. Scientific databases allow the integration of dissimilar data sets and allow data to be analysed in new ways, often across disciplines, making new types of scientific inquiry possible.

There are several advantages in developing and using scientific databases. First, databases lead to an overall improvement in data quality. The rise in user numbers increases the frequency of detecting and correcting problems in the data. A second advantage is the cost. It generally costs less to save than to recollect data. Also, in many cases uncontrollable factors (such as weather, population influences and ecosystem processes) make data recollection impossible, at any cost.

4.13 Patient reported feedback

In order to improve the quality of care, patient outcomes and to know how much the condition of the patient and its associated effects on them have improved the health services receive patient reported feedbacks. The health services receive this information by listening to the patients. However, just listening is not enough and the health services require special tools to collect patient reported feedback. Patient reported feedback can be formal or informal. One formal method is patient reported outcome measures (PROMS), which can be questionnaires for the patients that ask them about their health and quality of life before and after treatment. Informal feedback method includes reviews about the healthcare services on their website, blogs, any social media or online forum (“Using patient experience feedback to improve health services — CLAHRC,” n.d.).

4.14 Chemotherapy Regimen data

Chemotherapy regimen consists of single and multiple drug agents administered to patients in cyclical patterns known as treatment cycles. Physicians follow chemotherapy regimen protocols with these potentially toxic drugs. In order to determine what regimens are to be used in the current clinical practice or whether regimen variation changes efficacy and effectiveness is what the chemotherapy databases are used for (Syed and Das, 2015).

OncoRx is one such oncology-specific database which works around chemotherapy regimens and is used in anticancer treatment. Data provided by this database to clinicians includes pharmakokinetic parameters of the drugs and information regarding the chemotherapy regimens (Yap et al., 2010).

5 Utility of the solutions

Companies that are helping pharmaceutical industries to bring drugs into the market faster include Aetion AI technology, BERG, Clinerion. Aetion and Berg along with Andaman7, Concerto health, Cota healthcare, Deep lens, Exploris, Fitango health, Flatiron, Healthmyne and IBM Watson also take part in supporting the decision making processes for treatments of complex patients. Aetion for example, supports in decisions making for treatments of complex patients who are not represented in

RCTs and is also involved in improving patient outcomes at lower cost along with Berg and Fitango Health.

The companies that are involved in Clinical trial recruitment include Andaman7, BERG, Clinerion, Cognizant, Cota healthcare. Andaman7 for example, has created a PHR (Personal Health Record) to allow all patients to collect their health data on their smartphone (no medical data is stored in the cloud). This PHR is combined with a HIP - a Health Intermediation Platform to facilitate medical research that includes clinical trial recruitment process for the pharma companies, real world evidence/ data collection and decision support to help the providers to improve care for the patients. Data is only shared with the patient consent, in full respect of his privacy (GDPR and other regulations).

Berg, Clinerion, Cognizant and Cota healthcare are also involved in drug discovery and development. Clinerion uses AI technology for analysis of real world data that saves the clinical research industry months in time and millions in costs during drug development in reducing the need for study protocol amendments and in speeding up trial site selection and patient search and identification, i.e. finding enough patients within the target enrollment time. It increases the efficiency of the clinical trial process by enabling patient search and selection of sites with real and complete information (no randomness, or guesswork). Their Patient Network Explorer platform identifies suitable clinical study candidates for a given protocol through automated screening of electronic health records (EHRs), using real-time data. This enables optimization of study protocols, more efficient site selection and faster patient search and identification, leading to significantly more patients being enrolled.

6 Discussion

Syapse is a US based company, which is focused on the use of data science in oncology. It has a Research collaboration agreement (RCA) with the FDA Oncology Center for Excellence (OCE) focused on the use of real world evidence to support decision making in oncology. Data science use is evolving and Syapse along with OCE are now working with the stakeholders across the FDA by providing real world evidence to address the questions relating to testing and treatment patterns in patients, dosing and safety, and outcomes in oncology, with a focus on precision medicine. Syapse with the collaboration of FDA is investigating methods by using data science to derive real-world evidence from different sources like clinical data from EHRs and genomics data. The organizations will then utilize this data to characterize the regulatory suitability of real-world evidence (HealthITAnalytics, 2019).

The FDA and Syapse aims to inspect the real-world endpoints for tumors and malignancies, characterizing the utilization and the clinical impact of molecular testing. They will also deal with the patients outcomes and the adverse events in patients receiving precision medicine versus the clinical trial populations and incorporate these patient reported outcomes into real world evidence. This shows the efforts of the organizations to improve oncology care by using data science with real world evidence. The efforts of Syapse and FDA will help in improving the oncology care outcomes and the data collected using data science by these organizations has helped researchers in a variety of their oncology projects. Real world evidence collected and generated by using data science from EHRs and other data sources play an important role in the successful clinical trials. Syapse will engage oncologists in the use of real world evidence that they generate to inform decision making, which will improve the cancer treatments and modernize the clinical trials (HealthITAnalytics, 2019).

Data science is helping to accelerate the innovative cancer care by the use of real world evidence for personalized medicine and other solutions. With the advancements in the data science the information of real world evidence is being used to understand how the drugs that are being developed perform in real patient groups outside of clinical trials. RWE not only impacts clinical research but also clinical decision making by giving the oncologists a better understanding of the

therapies and treatments could affect the patient's health. RWE that is being generated is playing an important role in changing the regulatory approval processes for therapeutics in the development pipeline even before the drugs get to the market. This may increase the development and clinical research for the new drugs. By utilizing RWE the pharmaceutical companies can get the right drugs to the right patients much faster and more efficiently. A lot of the drugs on the market have been receiving extended approval based on the RWE.
The healthcare industry has faced challenges with clinical trials as they do not represent the whole patient population who get the treatments and are very time consuming and costly. It has now become impossible to rely on the information collected from clinical trials to support the regulatory decision making. The oncologists consider subjective as well as objective indicators to figure out the success of the treatments and it is not restricted to controlled settings (Doyle, 2019).

The use of Data science enables to use data from different data sources to enhance precision medicine use by the cancer care providers in personalized cancer care. Outside the use of clinical trials, the physicians rely on their experiences to determine whether or not a treatment would give the best outcome to a specific patient or not. The use of RWE comes in action here as by utilizing and organizing the deidentified patient data by using data science, RWE can give accurate information and put the pieces together for better understanding of real world clinical evidence in real time for the oncologists to act at the point of care and also reduce the costs. By the help of organized real world data the physicians can compare different patients' genetic makings to discover treatments that would work best. By using the right data to its maximum the oncologists can identify the treatment pathway faster and this may help them reduce the overall costs of the patients by removing the costly treatments which are ineffective in clinically similar patients (Doyle, 2019).

Data science is the future of Oncology care as the collection of data from different data sources like EHRs that collect huge amounts of important patient information from healthcare organizations, life sciences companies, physicians etc requires meaningful actions with this collected data. This is put into best use by real world evidence which helps in determining the best treatment pathways in oncology care. It helps the oncologists to view patient data in a clear manner and that they can make sure that every patient finds the right treatment pathway at the right time.

7 References:

Alamanou, @Marina T., 2019. Cancer and AI in a Single Frame [WWW Document]. Medium. URL https://towardsdatascience.com/cancer-and-ai-in-a-single-frame-bebd6b1b8e3 (accessed 8.1.19).

Barbosa, C.D., 2017. Challenges with Big Data in Oncology. J Orthop Oncol 02. https://doi.org/10.4172/2472-016X.1000112

Berger, M.L., Doban, V., 2014. Big data, advanced analytics and the future of comparative effectiveness research. J Comp Eff Res 3, 167–176. https://doi.org/10.2217/cer.14.2

Birnbaum, H.G., Cremieux, P.Y., Greenberg, P.E., LeLorier, J., Ostrander, J., Venditti, L., 1999. Using Healthcare Claims Data for Outcomes Research and Pharmacoeconomic Analyses: PharmacoEconomics 16, 1–8. https://doi.org/10.2165/00019053-199916010-00001

Bulik-Sullivan, B., Busby, J., Palmer, C.D., Davis, M.J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., Ojo, N.C., Caldwell, K., Abhyankar, J., Boucher, T., Hart, M.G., Makarov, V., Montpreville, V.T.D., Mercier, O., Chan, T.A., Scagliotti, G., Bironzo, P., Novello, S., Karachaliou, N., Rosell, R., Anderson, I., Gabrail, N., Hrom, J., Limvarapuss, C., Choquette, K., Spira, A., Rousseau, R., Voong, C., Rizvi, N.A., Fadel, E., Frattini, M., Jooss, K., Skoberne, M., Francis, J., Yelensky, R., 2018. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. Nat. Biotechnol. https://doi.org/10.1038/nbt.4313

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356, 183–186. https://doi.org/10.1126/science.aal4230

Cardoso, F., van't Veer, L.J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J.-Y., Brain, E., Causeret, S., DeLorenzi, M., Glas, A.M., Golfinopoulos, V., Goulioti, T., Knox, S., Matos, E., Meulemans, B., Neijenhuis, P.A., Nitz, U., Passalacqua, R., Ravdin, P., Rubio, I.T., Saghatchian, M., Smilde, T.J., Sotiriou, C., Stork, L., Straehle, C., Thomas, G., Thompson, A.M., van der Hoeven, J.M., Vuylsteke, P., Bernards, R., Tryfonidis, K., Rutgers, E., Piccart, M., MINDACT Investigators, 2016. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N. Engl. J. Med. 375, 717–729. https://doi.org/10.1056/NEJMoa1602253

Carr, C.T., Hayes, R.A., 2015. Social Media: Defining, Developing, and Divining. Atlantic Journal of Communication 23, 46–65. https://doi.org/10.1080/15456870.2015.972282

Chen, R., Snyder, M., 2013. Promise of Personalized Omics to Precision Medicine. Wiley Interdiscip Rev Syst Biol Med 5, 73–82. https://doi.org/10.1002/wsbm.1198

Cohen, I.G., Amarasingham, R., Shah, A., Xie, B., Lo, B., 2014. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood) 33, 1139–1147. https://doi.org/10.1377/hlthaff.2014.0048

Coker, E.A., Mitsopoulos, C., Tym, J.E., Komianou, A., Kannas, C., Di Micco, P., Villasclaras Fernandez, E., Ozer, B., Antolin, A.A., Workman, P., Al-Lazikani, B., 2019. canSAR: update to the cancer translational research and drug discovery knowledgebase. Nucleic Acids Res 47, D917–D922. https://doi.org/10.1093/nar/gky1129

Correa, T., Hinsley, A.W., de Zúñiga, H.G., 2010. Who interacts on the web?: The intersection of users' personality and social media use. Computers in Human Behavior 26, 247–253. https://doi.org/10.1016/j.chb.2009.09.003

Deng, J., Xing, L., 2019. Big Data in Radiation Oncology. CRC Press.

Doyle, M., 2019. Utilizing real-world evidence to improve oncology care. MedCity News. URL https://medcitynews.com/2019/04/utilizing-real-world-evidence-to-improve-oncology-care/ (accessed 9.28.19).

Drazen, J.M., 2002. Who owns the data in clinical trial? Sci Eng Ethics 8, 407–411.

EMR vs EHR – What is the Difference? | Health IT Buzz [WWW Document], n.d. URL https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference (accessed 8.27.19).

Fessele, K.L., 2018. The Rise of Big Data in Oncology. Seminars in Oncology Nursing, Technology in Cancer Care 34, 168–176. https://doi.org/10.1016/j.soncn.2018.03.008

Flatiron Health [WWW Document], n.d. . Flatiron Health. URL https://flatiron.com/ (accessed 10.28.19).

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., Maathuis, M.H., Moreau, Y., Murphy, S.A., Przytycka, T.M., Rebhan, M., Röst, H., Schuppert, A., Schwab, M., Spang, R., Stekhoven, D., Sun, J., Weber, A., Ziemek, D., Zupan, B., 2018. From hype to reality: data science enabling personalized medicine. BMC Medicine 16, 150. https://doi.org/10.1186/s12916-018-1122-7

Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M.P., Overington, J.P., Papadatos, G., Smit, I., Leach, A.R., 2017. The ChEMBL database in 2017. Nucleic Acids Res. 45, D945–D954. https://doi.org/10.1093/nar/gkw1074

Gupta, U.C., 2013. Informed consent in clinical research: Revisiting few concepts and areas. Perspect Clin Res 4, 26–32. https://doi.org/10.4103/2229-3485.106373

Hannaford, P.C., Owen-Smith, V., 1998. Using epidemiological data to guide clinical practice: review of studies on cardiovascular disease and use of combined oral contraceptives. BMJ 316, 984–987.

HealthITAnalytics, 2019. FDA to Advance the Use of Real-World Evidence in Cancer Care [WWW Document]. HealthITAnalytics. URL https://healthitanalytics.com/news/fda-to-advance-the-use-of-real-world-evidence-in-cancer-care (accessed 9.20.19).

Henricks, W.H., 2011. “Meaningful use” of electronic health records and its relevance to laboratories and pathologists. J Pathol Inform 2. https://doi.org/10.4103/2153-3539.76733

Hide, W., 2005. Genome Databases, in: John Wiley & Sons, Ltd (Ed.), ELS. John Wiley & Sons, Ltd, Chichester, UK, p. a0005314. https://doi.org/10.1038/npg.els.0005314

J Seabrook, H., Stromer, J., Shevkenek, C., Bharwani, A., de Grood, J., A Ghali, W., 2014. Medical applications: A database and characterization of apps in Apple iOS and Android platforms. BMC research notes 7, 573. https://doi.org/10.1186/1756-0500-7-573

Jensen, P.B., Jensen, L.J., Brunak, S., 2012. Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. 13, 395–405. https://doi.org/10.1038/nrg3208

Katal, A., Wazid, M., Goudar, R.H., 2013. Big data: Issues, challenges, tools and Good practices, in: 2013 Sixth International Conference on Contemporary Computing (IC3). Presented at the 2013 Sixth International Conference on Contemporary Computing (IC3), IEEE, Noida, India, pp. 404–409. https://doi.org/10.1109/IC3.2013.6612229

Kwon, O., Lee, N.Y., Shin, B., 2014. Data quality management, data usage experience and acquisition intention of big data analytics. Int J. Information Management 34, 387–394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539
Lee, C.H., Yoon, H.-J., 2017. Medical big data: promise and challenges. Kidney Res Clin Pract 36, 3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3
Mgudlwa, S., Iyamu, T., 2018. Integration of social media with healthcare big data for improved service delivery. SA Journal of Information Management 20, 8. https://doi.org/10.4102/sajim.v20i1.894
Mitsopoulos, C., Schierz, A.C., Workman, P., Al-Lazikani, B., 2015. Distinctive Behaviors of Druggable Proteins in Cellular Networks. PLOS Computational Biology 11, e1004597. https://doi.org/10.1371/journal.pcbi.1004597
Montouchet, C., Thomas, M., Anderson, J., Foster, S., n.d. The oncology data landscape in Europe: report 25.
Mullard, A., 2017. The drug-maker's guide to the galaxy. Nature 549, 445–447. https://doi.org/10.1038/549445a
Ni, D., Jin, L., Tu, W., 2017. Chapter 4 - Response to Early Warning Signals, in: Yang, W. (Ed.), Early Warning for Infectious Disease Outbreak. Academic Press, pp. 75–98. https://doi.org/10.1016/B978-0-12-812343-0.00004-7
Nimrod, G., Fischman, S., Austin, M., Herman, A., Keyes, F., Leiderman, O., Hargreaves, D., Strajbl, M., Breed, J., Klompus, S., Minton, K., Spooner, J., Buchanan, A., Vaughan, T.J., Ofran, Y., 2018. Computational Design of Epitope-Specific Functional Antibodies. Cell Rep 25, 2121-2131.e5. https://doi.org/10.1016/j.celrep.2018.10.081
Osong, A.B., Dekker, A., van Soest, J., 2019. Big data for better cancer care. Br J Hosp Med 80, 304–305. https://doi.org/10.12968/hmed.2019.80.6.304
Ozdalga, E., Ozdalga, A., Ahuja, N., 2012. The Smartphone in Medicine: A Review of Current and Potential Use Among Physicians and Students. J Med Internet Res 14. https://doi.org/10.2196/jmir.1994
Patel, M.N., Halling-Brown, M.D., Tym, J.E., Workman, P., Al-Lazikani, B., 2013. Objective assessment of cancer genes for drug discovery. Nat Rev Drug Discov 12, 35–50. https://doi.org/10.1038/nrd3913
Payne, K.F.B., Wharrad, H., Watts, K., 2012. Smartphone and medical related App use among medical students and junior doctors in the United Kingdom (UK): a regional survey. BMC Med Inform Decis Mak 12, 121. https://doi.org/10.1186/1472-6947-12-121
Petrov, C., n.d. Big Data Statistics 2019. Tech Jury. URL https://techjury.net/stats-about/big-data-statistics/ (accessed 10.29.19).
Pharmacy Drug Databases & Web Resources [WWW Document], n.d. . Study.com. URL https://study.com/academy/lesson/pharmacy-drug-databases-web-resources.html (accessed 8.15.19).
Rich, J., n.d. Library Guides: Data Resources in the Health Sciences: Clinical Data [WWW Document]. URL //guides.lib.uw.edu/hsl/data/findclin (accessed 8.26.19).
Sagiroglu, S., Sinanc, D., 2013. Big data: A review, in: 2013 International Conference on Collaboration Technologies and Systems (CTS). Presented at the 2013 International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47. https://doi.org/10.1109/CTS.2013.6567202
Sellwood, M.A., Ahmed, M., Segler, M.H., Brown, N., 2018. Artificial intelligence in drug discovery. Future Med Chem 10, 2025–2028. https://doi.org/10.4155/fmc-2018-0212
Silva, C., Fresco, P., Monteiro, J., Rama, A.C.R., 2013. Online drug databases: a new method to assess and compare inclusion of clinically relevant information. Int J Clin Pharm 35, 560–569. https://doi.org/10.1007/s11096-012-9746-8

Sobradillo, P., Pozo, F., Agustí, A., 2011. P4 medicine: the future around the corner. Arch. Bronconeumol. 47, 35–40. https://doi.org/10.1016/j.arbres.2010.09.009
Srkalovic, G., 2019. Genomics in Hematology and Oncology Practice. Acta Med Acad 48, 1–5. https://doi.org/10.5644/ama2006-124.237
Syed, H., Das, A.K., 2015. Identifying Chemotherapy Regimens in Electronic Health Record Data Using Interval-Encoded Sequence Alignment, in: Holmes, J.H., Bellazzi, R., Sacchi, L., Peek, N. (Eds.), Artificial Intelligence in Medicine, Lecture Notes in Computer Science. Springer International Publishing, pp. 143–147.
Taglang, G., Jackson, D., 2016. Use of "big data" in drug discovery and clinical trials. Gynecologic Oncology 141, 17–23. https://doi.org/10.1016/j.ygyno.2016.02.022
The future of cancer genomics, 2015. . Nat. Med. 21, 99. https://doi.org/10.1038/nm.3801
Trifiletti, D.M., Showalter, T.N., 2015. Big Data and Comparative Effectiveness Research in Radiation Oncology: Synergy and Accelerated Discovery. Front. Oncol. 5. https://doi.org/10.3389/fonc.2015.00274
Trivedi, H., Kling, H.M., Treece, T., Audeh, W., Srkalovic, G., 2019. Changing Landscape of Clinical-Genomic Oncology Practice. Acta Med Acad 48, 6–17. https://doi.org/10.5644/ama2006-124.238
Using patient experience feedback to improve health services — CLAHRC [WWW Document], n.d. URL https://www.clahrc-oxford.nihr.ac.uk/research/using-patient-experience-feedback-to-improve-health-services (accessed 11.4.19).
van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536. https://doi.org/10.1038/415530a
Ventola, C.L., 2014. Mobile Devices and Apps for Health Care Professionals: Uses and Benefits. P T 39, 356–364.
Vogenberg, F.R., Isaacson Barash, C., Pursel, M., 2010. Personalized medicine: part 1: evolution and development into theranostics. P T 35, 560–576.
Wang, Y., Bryant, S.H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B.A., Thiessen, P.A., He, S., Zhang, J., 2017. PubChem BioAssay: 2017 update. Nucleic Acids Res. 45, D955–D963. https://doi.org/10.1093/nar/gkw1118
Wedge, D.C., Gundem, G., Mitchell, T., Woodcock, D.J., Martincorena, I., Ghori, M., Zamora, J., Butler, A., Whitaker, H., Kote-Jarai, Z., Alexandrov, L.B., Van Loo, P., Massie, C.E., Dentro, S., Warren, A.Y., Verrill, C., Berney, D.M., Dennis, N., Merson, S., Hawkins, S., Howat, W., Lu, Y.-J., Lambert, A., Kay, J., Kremeyer, B., Karaszi, K., Luxton, H., Camacho, N., Marsden, L., Edwards, S., Matthews, L., Bo, V., Leongamornlert, D., McLaren, S., Ng, A., Yu, Y., Zhang, H., Dadaev, T., Thomas, S., Easton, D.F., Ahmed, M., Bancroft, E., Fisher, C., Livni, N., Nicol, D., Tavaré, S., Gill, P., Greenman, C., Khoo, V., Van As, N., Kumar, P., Ogden, C., Cahill, D., Thompson, A., Mayer, E., Rowe, E., Dudderidge, T., Gnanapragasam, V., Shah, N.C., Raine, K., Jones, D., Menzies, A., Stebbings, L., Teague, J., Hazell, S., Corbishley, C., CAMCAP Study Group, de Bono, J., Attard, G., Isaacs, W., Visakorpi, T., Fraser, M., Boutros, P.C., Bristow, R.G., Workman, P., Sander, C., TCGA Consortium, Hamdy, F.C., Futreal, A., McDermott, U., Al-Lazikani, B., Lynch, A.G., Bova, G.S., Foster, C.S., Brewer, D.S., Neal, D.E., Cooper, C.S., Eeles, R.A., 2018. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nat. Genet. 50, 682–692. https://doi.org/10.1038/s41588-018-0086-z
What is a Cancer Registry? [WWW Document], n.d. . SEER. URL https://seer.cancer.gov/registries/cancer_registry/index.html (accessed 8.26.19).

White, M.C., Babcock, F., Hayes, N.S., Mariotto, A.B., Wong, F.L., Kohler, B.A., Weir, H.K., 2017. The History and Use of Cancer Registry Data by Public Health Cancer Control Programs in the United States. Cancer 123, 4969–4976. https://doi.org/10.1002/cncr.30905

Workman, P., Al-Lazikani, B., 2013. Drugging cancer genomes. Nat Rev Drug Discov 12, 889–890. https://doi.org/10.1038/nrd4184

Workman, P., Antolin, A.A., Al-Lazikani, B., 2019. Transforming cancer drug discovery with Big Data and AI. Expert Opinion on Drug Discovery 1–7. https://doi.org/10.1080/17460441.2019.1637414

Yamada, I., Lopez, G., 2012. Wearable Sensing Systems for Healthcare Monitoring. Digest of Technical Papers - Symposium on VLSI Technology 5–10. https://doi.org/10.1109/VLSIT.2012.6242435

Yap, K.Y.-L., Chan, A., Chui, W.K., Chen, Y.Z., 2010. Cancer informatics for the clinician: An interaction database for chemotherapy regimens and antiepileptic drugs. Seizure 19, 59–67. https://doi.org/10.1016/j.seizure.2009.11.004

Yin, J., Lampert, A., Cameron, M., Robinson, B., Power, R., 2012. Using Social Media to Enhance Emergency Situation Awareness. IEEE Intelligent Systems 27, 52–59. https://doi.org/10.1109/MIS.2012.6

Young, S.D., 2014. Behavioral insights on big data: using social media for predicting biomedical outcomes. Trends Microbiol 22, 601–602. https://doi.org/10.1016/j.tim.2014.08.004

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L.D., Ferretti, V., 2019. The International Cancer Genome Consortium Data Portal. Nat. Biotechnol. 37, 367–369. https://doi.org/10.1038/s41587-019-0055-9