Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

[Kareem Hosny] Date

[TRIAL on FHIR]

By

[Kareem Hosny]
[Degree to be awarded: MPH]

[Applied Public Health Informatics]

_____

[Sheri Chernetsky Tejedor, MD, SFHM]
Committee Chair

_____

[Laurie Caydos, PhD]
Committee Member

_____

[Mark Conde]
Committee Member

[TRIAL on FHIR]

By

[Kareem Hosny]

[MD, MBBcH]
[Cairo University School of Medicine]
[2007]

Thesis Committee Chair: [Sheri Chernetsky Tejedor, MD, SFHM]

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
[Master of Public Health]
in [Applied Public Health Informatics]
[2017]

Abstract

TRIAL ON FHIR

By [Kareem Hosny]

- **Introduction:**

Every day scientists push the boundaries to discover new treatments. Treatments like antibiotics and vaccines that bring about profound changes in the world. But unfortunately many of these new treatments never pass the testing stage. The two main reasons for clinical trial failures are a steady decline in the number of patients recruited and failure to maintain patients in the trial. It is estimated that recruitment rates in clinical trials have dropped 20% since 2000, largely due to the fact that many patients lack knowledge about clinical trials that may be suited for their medical conditions.

- **The necessity:**

Patients with chronic diseases usually feel they are fighting an uphill battle against their depleting disease. Clinical trials provide patients with options that would otherwise not be part of their standard of care, and in doing so, they give patients a feeling of control over their disease and hope for the future. But the problems of recruitment and retention continue to plague researchers. In 2010 an estimated fifteen billion dollars was spent to enhance patient recruitment and continuation of clinical trials.

- **The proposal:**

We are proposing the development of a novel phone application and web-based software interconnected through a database that will match patients with cutting-edge experimental and investigational treatment options based on their medical conditions and proximity to clinical trial sites.

- **Methodology:**

We plan to build a platform and software algorithm assembled on: geographic information system (GIS) data mining, neural network machine learning (artificial intelligence), Natural Language Processing (NLP), artificial intelligence, and HL7 FHIR standards to match patients' medical records with clinical trial enrollment criteria.

- **Conclusion:**

We plan to create a platform that we hope will provide patients access to the most current and cutting-edge experimental and investigational treatment options and hopefully give back the control that their disease has taken from them. Our software will also help to increase the patient pool available to researchers, reduce the time spent on finding suitable candidates, and finally drive down the cost associated with conducting trials.

[TRIAL on FHIR]

By

[Kareem Hosny]

[MD, MBBcH]
[Cairo University School of Medicine]
[2007]

Thesis Committee Chair: [Sheri Chernetsky Tejedor, MD, SFHM]

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
[Master of Public Health]
in [Applied Public Health Informatics]
[2017]

TRIAL ON FHIR

(BULDING BRIDGES BETWEEN CURRENT PATIENTS AND FUTURE MEDICINE)

1. Introduction:

The evolution of medicine is rapid and exponential. Every day, scientists and research-physicians push the boundaries to discover new promising treatments for challenging diseases such as cancer, diabetes, and cardiovascular diseases. New treatments are only implemented as a new standard of care when all stages of research have been successfully completed. Stages of research include: laboratory bench studies, small and large animal model studies, transitional studies, human studies, and finally clinical trials. Clinical trial is the last and perhaps the most important stage that brings the entire research project together. Through clinical trials an idea becomes a reality; a chemical compound becomes a weapon against a disease.

Clinical trials are important for developing new management strategies for diseases, and they are also important in detecting new, cheaper, faster, and more accurate ways for diagnosing diseases. Additionally, they help identify risk factors of different diseases. In summary, they help scientists determine which part of their research hypothesis is working and which part is not (1).

Moreover, clinical trials are one of the main tools to shape the future of public health. By default, clinical trials deal with human health in a population instead of individually, as does traditional medicine. This public/population prospective of clinical trials is a main driver in improving public health. Evidence-based knowledge generated from clinical trials creates the cornerstone for therapeutic guidelines in mass populations. They steered the implementation of new standards of care by introducing new effective drugs

and devices that improve survival, reduce mortality, and prevent disease complications. They are also responsible for minimizing the risk of ineffective and/or unsafe medical interventions. It has been proven that many aspects of public health, especially in the field of epidemiology, have evolved due to the conduction of clinical trials. They provided significant information regarding disease trends, risk factors, treatment outcomes, health care costs, and public health impacts of different interventions (2).

In addition, researchers have investigated the public return on the investment of clinical trials. For instance, all clinical trials funded by USD National Institute for Neurological Disorders and Stroke before 2000 were evaluated for the efficacy of trial results on medical care and health. On the public health level, 21% of the trials showed measurable improvement on health and 14% showed significant medical cost reduction to society. On the long-term investment level, clinical trials were responsible for 470,000 quality-adjusted life years at a total cost of 3.6 billion dollars during a 10 year period. In conclusion of this study, the net benefit to society for this 10 year period was 15.2 billion dollars from clinical trials (2). The study saved a net total of $15.2 billion over the same 10 years from clinical trials, freeing up further capital for other studies (3).

To conduct clinical trials, researchers need a large pool of patients to generate meaningful results and in the past decade, public interest has increased in medical research, in particular toward clinical trials. Patients are generally more interested in participation due to different benefits and motivations.

Below are the most common reasons for participation in clinical trials (4,5):

1) Access to new promising treatment options
2) Enhanced medical care with free follow up visits and clinical monitoring

3) Altruism and selflessness

4) Financial incentives

Recent polls show that 80 % of Americans have heard about clinical trials and 74% of Americans believe that their health has been improved due to innovation in medical research.

Even though there is an increase awareness about clinical trials, enrollment rates for pharmaceutical studies have dropped 20% since 2000, as only around 10% of Americans reported participating in clinical trials, and to make matters worse only 22% of the patients reported that their medical provider discussed clinical trials for their conditions (6,7). All of this leads to a delay in implementation of around 85% of clinical trials.

Clearly there is a mismatch between public awareness and clinical trial participation. The reasons for this include patient location, patient awareness about a particular trial suited for their medical condition, vague understanding of the requirements, and risks and benefits associated with clinical trials. Furthermore, many patients hold false beliefs that they will not qualify for the trials, confusion about the insurance coverage, and the misunderstanding that they will end up paying for participation. Barriers like incompatibility with medical standards of care, inflexibility with schedules, and a lack of both financial and non-financial incentives can lead to failure of enrollment of suitable patients in clinical trials (5,8-10). It is no wonder that more than 60% of clinical trials reported difficulty in recruiting a significant number of patients (5,8,10).

The mismatch between public awareness and clinical trials not only puts a strain on the type of research that can be conducted but it also leads to a substantial increase in research costs. In 2010 pharmaceutical research companies spent $46.4 billion on clinical

trials, almost double the amount spent over the last decade. Of that $46.4 billion, almost a third ($15 billion) was spent on patient recruitment in the trials as well as maintaining them from dropping off (11). The average cost varies between $5,500–ß7,600 per single patient. Yet there is continued delay in recruitment that leads to drug development failure. The new treatment approval rate has been flat since 1990 despite the huge inflation of research spending in the last decade (11,12).

Moreover, the majority of patients do not receive their medical care in academic institutions with established research opportunities for clinical trials. They may receive medical care in community hospitals or doctor's offices with less experience in clinical trials than their counterparts may. In the past, community hospitals were not particularly involved in clinical research, but recently clinical trials have been getting particular attention in community hospitals, whether it is due to the vast reach of the internet as the general population becomes aware of the potential treatment options, or due to the vigorous implementation of clinical trials in community hospitals after the initiation of CCOP's (Community Clinical Oncology Programs) and then the establishment of CTSU (Clinical Trial Service Unit) in the 1980s.

Many community hospitals have realized the need for developing strong medical research programs in order to keep pace in the ever-evolving healthcare market. However, promoting information regarding these programs to the general population has many pitfalls. Patients still have issues in allocating suitable research sites for their medical conditions.

There is an urgent need to create essential tools to close the gap between patients' awareness about clinical trials and participation in the trials. Sufficient recruitment of

volunteer subjects is the golden cornerstone for decreasing research expenditures, maintaining research project schedules, and acquiring sufficient data that will provide high-yield data with strong statistical analysis needed to get the treatments approved by FDA. Researchers are working tirelessly to come up with new ways to enhance recruitment and continuation of trials. Below are the five most common strategies being deployed to combat decline in recruitment (10):

1) Newsletters/ mailers/ flyers (to both medical care providers and/or patients)
2) Follow-up recruitment visits to emphasize different available trials
3) Follow-up phone calls to care providers to explain different available options of clinical trails
4) Posters in medical care centers and clinics
5) Protocol amendment and changing in inclusion/exclusion criteria to expand criteria for participation
6) Presentations for interest groups

While these strategies tend to improve enrollment, recruitment remains an ongoing struggle. Great minds must come together to create new theoretical and practical strategies, new tools that use cutting edge technology to enhance enrollment rates. Additionally, these new tools must be accessible to the patients as well as the researchers.

One such tool currently being deployed includes an online database that matches patients through synchronizing a few key features of trials and patient records. However, this synchronization between different vendors, entities and data sources creates a huge problem with interoperability. Additionally, these new methods do not address a key issue with recruitment: patient geographical location in terms of the clinical site. As

mentioned earlier, the majority of patients receive their medical care in community hospitals, and it is not feasible for them to travel long distances to another city in order to be enrolled in a clinical trial. Based on the fact sheet by American Health Association, there are 5564 hospitals in USA, with the majority—4,862—community hospitals. The geographic distance between patients and clinical trial locations not only impacts the enrollment rates of clinical trials but it also affects the attrition rate, the process of losing the patient after recruitment. It has been documented that patients who live far away from research hubs are less involved in enrollment with high attrition rates (10,13,14).

To our knowledge there is not a single approach that implements GIS data mining with machine learning models while using data elements standards such as HL7 FHIR standards to bring patients and researchers together (8).

1.1. Why HL7-FHIR

FHIR (pronounced "fire") is a novel set of standards and resources that allows inquiry of Electronic Healthcare Record (EHR) systems to retrieve and insert patient's health and demographic data into a database. It is a budding project from the international standards Health Level 7 (HL7) that is used to transfer clinical and administrative data between software applications among different healthcare vendors. FHIR was developed to allow for a more customized experience by end-users while using HL7 standards. This 'user-friendly' model improves the process of implementation of these standards by different vendors.

FHIR consists of hundreds of tiny sets of core information resources that can be bundled up and deployed to various database systems using IT platforms such as SMART Substitutable Medical Apps, Reusable Technology) on FHIR. This process of aggregation

and deployment can be used to answer various queries. FHIR utilizes similar technology that drives web platforms like Facebook and Twitter. This is why it is best suited to develop patient-centric/patient-friendly applications. It is also compatible with different operating systems for different devices including PCs, tablets or smart phones. FHIR is an open source, i.e. free to use by patients, physicians, researchers and pharmaceutical companies and can be easily used by IT programmers, patients, and scientists (15).

Furthermore, FHIR features can be used to develop a standard Application Programming Interface (API) for:

1) Assessing the eligibility criteria for different pharmaceutical trials
2) Supporting patient recruitment into these trials based on their medical history
3) Harvesting data (both structured and unstructured) from EHRs to support clinical studies
4) Instant stratification of data through applying and answering different queries.

The potential benefits to harvest, retrieve, process and resend data using FHIR standards are very promising. As an open source, it opens a path to integrate and analyze high quality digital healthcare data in an instant, creating a revolution in the world of pharmaceutical trials. This new approach provides a comprehensive and innovative solution in the advancement of pharmaceutical trials while keeping the focus on the patient. It also gives patients the power to make medical decisions that are in their best interest.

Additionally, the use of FHIR in our algorithm will help achieve two main overarching goals identified by the Meaningful Use Policy for electronic exchange of

health information; improving patients' health and decreasing cost (3). FHIR will meet these by providing comprehensive solutions and resources through the following:

1) Satisfying patients demand for flexible access to their own health information: FHIR provides an application program interface (API) that allows flexible end-user friendly access to patients' own health information.
2) Offer health care providers faster, interoperable access to patient records: The different resources (data models) that FHIR provides allow easier, faster Interoperability between health care providers.
3) Reduce errors in recording vital information in patients chart: The use of electronic exchange of information through FHIR instead of the traditional Fax/mail has the potential of decreasing human errors associated with recording data (i.e. date recording into patient's medical records) and translating it from source documents.
4) Provides research data of unprecedented power to transform health care: FHIR standards will provide substantial data sources that can be further utilized to power clinical care, public health and biomedical research.
5) Facilitates communication among health care providers and patients regarding clinical trials: It will improve data exchange through EHRs that have been pushed very hard by Health Information Technology (David Blumenthal) and Centers of Medicare & Medicaid Services, CMS (Marilyn Tavenner). This push will be a great opportunity for FHIR to thrive (16).

Another great opportunity for using FHIR is the uprising movement to integrate FHIR with major electronic health record (EHR) industry partners. This includes the following:

- Accenture

- Athenahealth
- Beth Israel Deaconess Medical Center
- Cerner
- Epic
- Intermountain Healthcare
- Mayo Clinic
- MEDITECH
- McKesson
- Partners HealthCare System
- SMART at the Boston Children's Hospital Computational Health Informatics Program
- The Advisory Board Company
- Surescripts

These industry partners have signed and agreed to support FHIR through the Argonaut Project furthering database promotion and usage. John Halamka (CIO of Beth Israel and Deaconess Medical Center) was quoted saying, "Imagine Epic and Cerner and Athena had a hook on their EHRs." This "hook" or application programming interface (API) is the key element, and what can enable EHRs to communicate interactively, and in a standardized way, with apps and eventually even other EHRs. Through this "hook", TRIAL on FHIR will be able to access patients' medical information and use different standards and resources provided through FHIR API to retrieve the required information to match the patient with the suitable clinical trials.

However, even with this initiative, EMR vendors are facing resistance in an agreement to support FHIR standards. After all Healthcare is a business enterprise, with many EMR vendors and exclusive partnerships that will not be willing to implement FHIR as their standard resources. This is an anticipated major challenge that we are expecting to face Mitigation of this challenge will be approached by using different innovative technology layers in our software, especially the NLP artificial inelegant component as will be discussed later.

To provide further clarification, illustration of some of the FHIR resources that we will use in building the application might be useful. Trial on FHIR app will support the current DSTU 2 Final (1.0.2) version of the FHIR standards. All API access is running over HTTPs. All data is sent and received as JSON or XML format. By implementing these specific resources (illustrated below), we are simplifying the level of complexity of the NLP layer of the App.

- **Condition:**

Description: "The Condition resource is used to record details about a patient's problems, diagnoses, or other health matters that are concerning. It is common to capture Conditions during the encounter or visit. This is usually captured both as initial suspected problems and confirmed or refuted problems or diagnoses at the time of discharge. This resource may be referenced by other resources as a "reason" for an action or order such as the reason for ordering a medication or procedure."

The following parameters are fulfilled if available:

- ID
- Patient

- ➢ Patient encounter when first recorded (only applies to diagnoses)
- ➢ Who recorded the condition
- ➢ Date recorded
- ➢ Condition code
- ➢ Status
- ➢ Category
- ➢ Verification status
- ➢ Onset (Date Time)
- ➢ Resolved (either Boolean or DateTime) (only applies to problems)
- ➢ Severity
- ➢ Comment/Note

- This resource can be used to match with different required medical conditions illustrated in the recruitment criteria of the clinical trials
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Information will be acquired from: Condition.Category, Condition.Clinicalstatus and Condition.Severity
- Request example:

```
GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/Condition?patient=4342012
```

- **Observation:**

Description: “The Observation resource provides measurements or simple assertions about a patient. Observations are crucial to supporting diagnoses, monitoring progress, and

establishing baselines or trends. Most observations are simply a name and value or result but some observations, such as blood pressure, group other observations together logically. Examples of common observations are: Laboratory Results (blood sugar, hemoglobin), Vital signs (temperature, blood pressure), Personal Characteristics (height, weight), and Social History (tobacco/alcohol use, employment status). Pathology reports, radiology reports, and other textual reports should be represented by the DiagnosticReport resource"

The following parameters are fulfilled if available:

- Observation id
- Status
- Category (laboratory, social history)
- Patient
- Patient encounter
- Effective date/time (collection date/time for laboratory)
- Issue date/time (date/time observation made available, entered, verified)
- Observation (name or text)
- Observation value or result
- For Observations with ValueQuantity
- Quantity comparator (<, <=, >, >=)
- Quantity units
- Interpretation (abnormal flagging)
- Reference range
- Comments
- Related Observations

- ➢ Components (e.g.: systolic and diastolic for blood pressure)
  - ➢ This resource **is the main** FHIR resource that we are planning to use. It will contain the majority of clinical measurements that are relevant to the patient medical conditions and their correlation to the recruitment criteria of the targeted clinical trials. . Observations are crucial to supporting diagnoses, monitoring progress, and establishing baselines or trends for participants in the targeted clinical trials
  - ➢ The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
  - ➢ Information will be acquired from: Observation.category, Observation.Code, Observation.Interpretation, Observation.Component.Code
  - ➢ Request example:

```
GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/Observation?patient=3998008
```

- **Device:**

  Description: “The Device resource implementation is currently limited to devices implanted in a patient such as a pacemaker or insulin pump. As such, this resource currently exposes the known implant history for a patient at this site”

  The following parameters are fulfilled if available:

  - ➢ Device id
  - ➢ Type
  - ➢ Manufacturer

- ➢ Model
- ➢ Manufacture date
- ➢ Expiration date
- ➢ UDI (FDA Unique Device Identifier)
- ➢ Serial number
- ➢ Lot number
- ➢ Owner (Facility)
- ➢ Patient
- ➢ This resource is essential if the existence of a medical device is important as an inclusion or exclusion criteria for the clinical trial
- ➢ The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- ➢ Information will be acquired from: Device.Identifier.Type, Device.Type
- ➢ Request example:

```
GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/Device?patient=4478007
```

- **Goal:**

  Description: “Codes for grouping and sorting goals”

  The following parameters are fulfilled if available:

  - ➢ ID
  - ➢ Patient
  - ➢ Date recorded

- ➢ Subject: Patient
- ➢ Targetdate

- This resource can be used to match with different anticipated goals and milestones illustrated in the recruitment criteria of the clinical trials
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Request example:

```
GET /Goal?:parameters
```

- **DiagnosticReport**

  Description: “Codes for diagnostic service sections”

  The following parameters are fulfilled if available:

  - ➢ ID
  - ➢ Patient
  - ➢ Subject: Patient
  - ➢ Date
  - ➢ _count

- This resource can be used to match with different clinical values in laboratory, radiology and pathology reports illustrated in the recruitment criteria of the clinical trials
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Request example:

```
GET /DiagnosticReport?:parameters
```

- **AllergyIntolerance:**

  Description: “The AllergyIntolerance resource provides the clinical assessment of a patient’s allergy or intolerance when exposed to a specific substance or class of substance including information about the adverse reaction. Substances include, but are not limited to, medications, foods, environment (such as plants and animals), and insect bites. The Allergy/Intolerance list exists as a patient safety tool for clinical decision support when ordering medications and nutrition or guiding clinical treatments. This resource does NOT include adverse reactions or adverse events which are expected for the circumstance such as an overdose or drug/drug interaction or an error/failure in the clinical process”

  The following parameters are fulfilled if available:

  - ➢ Allergy id
  - ➢ Allergy substance
  - ➢ Status
  - ➢ Patient with allergy/intolerance
  - ➢ Date of allergy onset
  - ➢ Date/Time recorded
  - ➢ Who recorded
  - ➢ Who reported
  - ➢ Criticality/potential harm
  - ➢ Category (medication, food, environment)
  - ➢ Adverse reaction (clinical symptoms/manifestation)

- ➢ Comment

- This resource is essential to determine if the presence or absence of a certain allergy or any substance intolerance will determine the enrollment status of the patient in the targeted clinical trial.
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Information will be acquired from: AllergyIntolerance.substance, AllergyIntolerance.reaction.manifestation
- Request example:

```
GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/AllergyIntolerance?patient=4342010
```

- **Procedure:**

  Description: "The Procedure resource returns current and historical procedures performed on a patient. The current implementation doesn't include surgical or radiologic procedures."

  The following parameters are fulfilled if available:

  - ➢ Procedure id
  - ➢ Patient
  - ➢ Status (completed, entered-in-error)
  - ➢ Procedure
  - ➢ Reason performed
  - ➢ Who performed
  - ➢ Date performed

- ➢ Patient encounter
- ➢ Notes

- This resource is essential to determine if the presence or absence of a procedure will determine the enrollment status of the patient in the targeted clinical trial.
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Information will be acquired from: Procedure.code, Procedure.performer.role
- Request example:

```
GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/Procedure?patient=1316020
```

- **Binary:**

  Description: "The Binary resource can contain any clinical content such as text, images, and PDFs. This resource is important for outside records and diagnostic reports."

  The following parameters are fulfilled if available:

  - ➢ Binary id
  - ➢ Content type
  - ➢ Content

- This resource is of an extremely important value in cases we are querying medical records in different non-structural formats.
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Information will be acquired from: Procedure.code, Procedure.performer.role
- Request example:

- GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/Binary/TR-5927259

- **Patient:**

Description: “The Patient resource provides general demographic information about a person receiving health care services from a specific organization. Common demographic fields include patient id, patient name, gender, date of birth, address, phone, primary language and marital status”

The following parameters are fulfilled if available:

- ➢ Patient name
- ➢ Patient id
- ➢ Extensions including birth time, birth sex, ethnicity, and race
- ➢ Medical Record number (MRN)
- ➢ Phone (email is not supported at this time)
- ➢ Contact person (guardian, parent or emergency)
- ➢ Gender (administrative)
- ➢ Date of Birth
- ➢ Address
- ➢ Communication (preferred language)
- ➢ Marital status

- This resource is of an extremely important value in the matching process. First it is important for patient identification proposes. Second, it is important in the matching

process in the GIS system for geographic matching to the closest active center of the clinical trial.

- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Information will be acquired from: Patient.identifier.type, Patient.contact.relationship
- Request example:

- GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/Patient?_id=4342009

- **Immunization:**

Description: "The Immunization resource includes the view of current and historical administration of vaccinations to a patient in all healthcare settings. This resource contains the functionality to query a patient's immunization history"

The following parameters are fulfilled if available:

- ➢ Id
- ➢ Vaccine administered
- ➢ Administration date/time
- ➢ Status
- ➢ Patient
- ➢ Was not given
- ➢ Reason not given
- ➢ Patient encounter
- ➢ Who administered
- ➢ Administration site

- ➢ Administration route
- ➢ Vaccine manufacturer
- ➢ Lot number
- ➢ Expiration date
- ➢ Dose

- This resource is essential to determine if the presence or absence of certain immunization as well as the date of the immunization will determine the enrollment status of the patient in the targeted clinical trial
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Information will be acquired from: Immunization.vaccineCode, Immunization.site, Immunization.route, Immunization.explanation.reasonNotGiven
- Request example:

- GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/Immunization?patient=4478007

- **MedicationOrder:**

Description: "The Medication Order resource provides orders for all medications along with administration instructions for a patient in both the inpatient and outpatient setting (orders/prescriptions filled by a pharmacy and discharge medication orders"

The following parameters are fulfilled if available:

- ➢ Medication order id
- ➢ Date/Time order written
- ➢ Status

- ➢ Date/Time order stopped or to stop
- ➢ Patient
- ➢ Prescriber
- ➢ Patient encounter
- ➢ Medication
- ➢ Details of medication taken:
- ➢ Dosage Instructions:
- ➢ Patient friendly dosage display
- ➢ Additional instructions
- ➢ Time period and frequency
- ➢ As needed (such as PRN for pain)
- ➢ Route
- ➢ Dosage
- ➢ Dispensing details:
- ➢ Validity period
- ➢ Number of refills
- ➢ Amount of medication to supply/dispense

- This resource is essential to determine if the presence or absence of certain medication, the dose of the medication and the date of administration will determine the enrollment status of the patient in the targeted clinical trial
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM

- Information will be acquired from: MedicationOrder.reasonCodeableConcept, MedicationOrder.medicationCodeableConcept, MedicationOrder.dosageInstruction.additionalInstructions, MedicationOrder.dosageInstruction.timing.code, MedicationOrder.dosageInstruction.asNeeded[x], MedicationOrder.dosageInstruction.siteCodeableConcept, Medication.product.form, Medication.code
- Request example:

- GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/MedicationOrder?patient=4342010

- **MedicationStatement:**

Description: "The Medication Statement resource provides a snapshot in time of known medications taken by the patient now or in the past reported by either the patient, significant other or a provider. Future orders are not returned. Documented historical/past/home medications are commonly captured when taking the patient's medical history. Prescriptions without documented compliance are intended, since we may not know if the patient is actively taking the medication or has filled the prescription. Medications are assumed to be taken unless documented otherwise"

The following parameters are fulfilled if available:

- ➢ Patient
- ➢ Source of information
- ➢ Date/Time recorded
- ➢ Status

- ➢ Was or Was Not Taken
- ➢ Date/Time started and ended
- ➢ Medication
- ➢ Reference to source MedicationOrder, if applicable
- ➢ Category Extension
- ➢ Details of medication taken:
- ➢ Dosage
- ➢ Route
- ➢ Frequency
- ➢ Quantity
- ➢ Reason for use
- ➢ Site
- ➢ Rate
- ➢ Order comments/special instructions
- ➢ Patient friendly display Extension

- This resource is essential to determine if the presence or absence of certain medication, the dose of the medication and the date of administration will determine the enrollment status of the patient in the targeted clinical trial
- The Terminology binding is supported by SNOMED CT, ICD-9-CM and ICD-10-CM
- Information will be acquired from: MedicationStatement.reasonForUse[x], MedicationStatement.medication[x], MedicationStatement.dosage.timing.code, MedicationStatement.dosage.asNeeded[x],

MedicationStatement.dosage.siteCodeableConcept,

MedicationStatement.dosage.route, Medication.code, Medication.product.form

- Request example:

- GET https://fhir-open.TRIALonFHIR.com/dstu2/0b8a0111-e8e6-4c26-a91c-5069cbc6b1ca/MedicationStatement?patient=4342010

On the other hand, we do not believe that TRIAL on FHIR app will need the following FHIR resources at least in the beginning:

- **Overview:** we will extract different data values from other resources. We are anticipating that information provided by overview will be of low yield.
- **Authorization:** we don't expect that the need for different level of personal authorizations will be vital for matching with clinical trials recruitment criteria.
- **Conformance:** We don't expect the need for capabilities and configurations in the matching process with clinical trials recruitment criteria.
- **OperationDefinition:** We don't expect that the need for documentation of continuity of care will help in matching with clinical trials recruitment criteria.
- **CarePlan:** We don't expect that the need for care plan verifications will be vital for matching with clinical trials recruitment criteria
- **Encounter:** We don't expect that administrative information will be vital for matching with clinical trials recruitment criteria
- **Contract:** We don't expect that contracts and personals what are authorized to view patient's information will be vital for matching with clinical trials recruitment criteria

- **Person:** We don't expect that individuals and contacts providing mechanism for person resources across a medical care facility will be vital for matching with clinical trials recruitment criteria
- **Practitioner:** We don't expect that information regarding individuals who are formally involved in the care of a patient will be vital for matching with clinical trials recruitment criteria
- **RelatedPerson:** We don't expect that information regarding individuals who are not-formally involved in the care of a patient will be vital for matching with clinical trials recruitment criteria
- **DocumentReference:** We don't expect references for documents especially continuity of care documents (CCD) will be vital for matching with clinical trials recruitment criteria
- **Appointment:** We don't expect that information on different types of clinical appointments and when they were created will be vital for matching with clinical trials recruitment criteria

1.2. The problem statement:

Many patients with end stage chronic diseases, especially cancer and cardiovascular patients believe they are fighting a losing battle. These patients are looking for an additional option that can potentially help improve their lives and give them hope toward their future. This new option can only come from new cutting edge advances implemented through clinical trials. But even if clinical trials provide an extra-option for patients, many patients do not know about specific trials that meet their needs. Additionally, patients cannot always locate the closest or the most convenient research

site available. On the other hand, researchers are in a continuous struggle to find suitable candidates for their trials. And even after enrollment, researchers struggle to maintain candidates during the follow up phase of the clinical trials. Finally, Public-health decision makers working towards improving the future of medicine are always searching for new and more convenient ways to develop new treatments for different diseases.
But the problem with patient recruitment remains an issue and new tools and techniques must be developed to increase patient recruitment.

1.3. The proposal statement:

In order to fill the void between public awareness and clinical trial recruitment, we propose the development of a mobile application built on GIS allocated data and machine learning behavior to match patients based on their medical conditions with appropriate clinical trials in close proximity to the patient. Our mobile application will apply a similar concept to that of UBER's platform in terms of matching driver' availability with riders desire to travel. The application will be built on GIS allocated data, natural language processing (NLP) data mining, machine learning behavior and will implement the Health Level 7- Fast Healthcare Interoperability Resources (HL7 FHIR) standard.

The application will allow the patient or designated medical provider (e.g. primary care physician, PCP) to enter the patient information in the application. The application will also give access to the researchers and pharmaceutical trial teams to post general information about the clinical trial and any requirements including inclusion/exclusion criteria. The software will then match the patient with the suitable clinical trial that focuses on his/her medical condition located in close proximity to the patient. Upon a

successful match, researchers will be given access to the patient's EHR data but only after consent has been given by the patient in accordance with the Health Insurance Portability and Accountability Act (HIPAA) guidelines.

2. Methodology:

2.1. Development process

- **Literature analysis:**

In order to explore appropriate concepts for the development of our model, we conducted a thorough literature review to determine the extent of various tools currently in practice to recruit and retain patients in clinical trials. The scientific databases used to explore current options included: PubMed, Web of Science Core Collection, Scopus, Library of Congress and Google Scholar. We also conducted our electronic searches in various IT, medical and informatics blogs through web-based search engines such as google.com and bing.com. We further investigated the reference lists in each original manuscript for more in depth knowledge about the discussed concepts and identify any new potential references.

- **Search Terms:**

We conducted our electronic search on the previous mentioned web-based scientific searching engines by searching for the following terms: "HL7", "FHIR", " standards solution for interoperability", "machine learning", "neural network", "deep learning", "supervised machine learning", "unsupervised machine learning", "GIS", " geographic data mining", "text data mining", "NLP", Natural Language Processing", "Clinical trials", "Pharmaceutical trials", "Enrollment rate in clinical trials", "Participation in clinical trials", " expenditure of clinical trials", " IT solution for clinical trials", "Public

knowledge about clinical trials", " "programming languages", " Clinical trials matching service", "finding suitable clinical trials", " registration for clinical trials", " enrolling in clinical trials" and " searching suitable clinical trials for me".

- **Characteristics of studies included for research:**

We generalized our search to include all scientific journals in the field of medicine, public health, health bioinformatics and IT technology due to the interdisciplinary characteristic of our project and these fields. There were no limits to publication dates for our search and there were no specific publication country or location.

- **Eligibility criteria of the literature exploration:**

Articles with the following criteria were eligible for our literature review:

1) Biographies, books, documents, blogs, clinical trials, dataset, dictionary, meta-analysis article, original journal manuscripts, official government publications, practice guideline, review article, systemic reviews, vendors websites and webcasts.
2) Articles and publications written in English
3) Articles and publications which included one or more of the previous mentioned search terms
4) Article and publications that contained references and citations to validate their work
5) Articles with limited accessibility to the full text were excluded
6) Articles conducted in basic research or in animal species were excluded

- **Identifying current and future initiatives**

Finally we tried to identify current vendors or initiatives targeting the problem of matching patients with clinical trials through the web service. To the best of our knowledge there was no vendor or initiative that proposed development of a machine-learning model using NLP, GIS text mining and Neural Network learning combined with HL7 FHIR standards.

3. Project outcome:

3.1. Search results:

Our comprehensive literature review identified new approaches for using HL7 FHIR in medical IT, best tools for the use of NLP, text mining and machine learning in the medical field. We also identified current approaches being utilized by different vendors for matching clinical trials with suitable patients. These vendors provide matching services by one of the following approaches(8):

1) Patient registration vendors: These vendors acquire key information including patients' demographics and medical conditions by requiring patients to fill out a profile and answering certain "schema" surveys. Then they compare these answers with the key features of the clinical trials available in the database. Examples include: "ResearchMatch", and "PatientsLikeMe"
2) Database search engine vendors: these are stand-alone vendors that do not require registration from the patients. They work by providing a database collection for patients to search trials in a certain location. These are not specific to a certain disease. Examples include: "CureLauncher" and "My Clinical Trial Locator".

3) Disease-Specific vendors: these are similar to the database search engine vendors but specialize in certain diseases and conditions. Examples include “Smart Patients” for cancer trials and “Fox Trail Finder” for Parkinson’s disease trials.
4) Vendors utilizing machine learning models: these are the most current method for matching patients with clinical trials or services. These work through the use of an artificial intelligence model to harvest information from clinical trials while at the same time harvesting patient information made available to them by the patients, the machine then synchronizes the gathered information to match patients with a trial. Examples include “IBM Watson for clinical trials” and “ Mendel.ai”

Lastly and most importantly, our search did not reveal an approach that addresses the key issue of a patients’ geographical location in recruitment. Current vendors either search a generic term in patient record to match clinical trials in a certain area or the most advanced machine learning systems do not utilize the geographical location at all.

The conceptual design of the model

There is an essential need to use universal IT standards to allow for appropriate synchronization, and develop an approach which implements GIS data mining, deep machine learning models and utilize data formats and/or data elements standards such as HL7 FHIR standards to increase recruitment rates. Additionally, new approach must be user friendly to all parties involved. Current matching services are not focusing on developing a simple, easy to use mobile platform that can be utilized by patient as well as the investigator. Therefore, we are proposing the development of a machine learning model that can process inclusion and exclusion criteria for clinical trials, and employs

these criteria's to the patient medical records while at the same time analyze and compare between clinical trial site location and patient's geographical location.

In order to match patients with trials the algorithm has to identify different synonyms for medical terminology in patient's records along with the clinical trial information, and at the same time match patients location based on city, state or zip code to the closest clinical site. Upon analysis of the data, our software will create a probability of a potential match between patients and the available clinical trials in patient's area.

The conceptual design of the application will be based on the following core model:

- **Two user interfaces:**

a. Participant interface:

In this interface, patients interested in clinical trial participation or their medical provider, or any medical provider associated with the patient will be able to search for the potentially beneficial clinical trial for the patients. The patient or his/her designated medical care provider will then be able to enter the patient information on the phone application. The application will then turn around and match the patient with the suitable clinical trial that focuses on his/her medical condition in the closest geographic location possible.

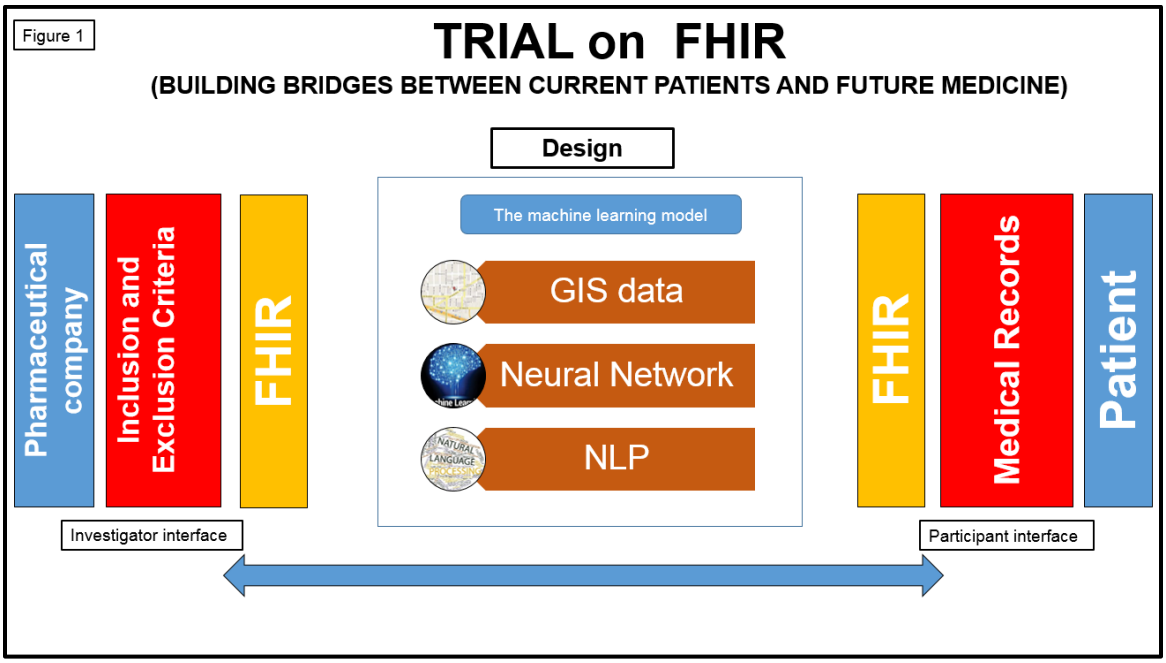
b. Investigator interface:

In this interface, any pharmaceutical investigator company, research clinician or principle investigator will be able to search for suitable candidates for their clinical trials. This model will expedite and improve the screening process, which in the past has been a bottleneck stage for many clinical trials.

- **4 components of the model algorithm:**

1) **GIS data:** geographic data mining from GIS data
2) **Machine learning:** using neural network model
3) **NLP:** text data mining using natural language processing
4) **HL7 FHIR standards**

Figure (1) illustrates the conceptual model of the application:



3.1.1. GIS data allocation (Geographic data mining)

GIS data mining will help bridge the gap between patients' needs and medical services provided. A lot of patients do not receive the appropriate medical care for their conditions because they simply live too far from the appropriate medical care facility. This gap widens when it comes to experimental treatments like pharmaceutical trials. We believe that GIS data mining is the key to increasing recruitment rate by matching the clinical trial location with patient's location

GIS is an IT information schema designed to implement IT solutions to the requested queries on spatial data. Encoding different variables on a geographic map and processing these variables interactions is how many GIS applications work. These applications allow users to generate interactive queries of data on spatial dimensions(17).

Many GIS applications implement the use of GIS for matching interactions between users' queries in bidirectional API platforms. Of those, UBER and Airbnb applications are among the most common applications that use GIS data. The Uber application visualizes millions of geodynamic data points to match drivers with customers. Airbnb operates by allowing hosts to post their own homes and apartments as a potential place to stay for guests, while allowing guests to find a suitable place to stay based on their address/zip code preference. Even for a huge company like Airbnb, its platform utilizes a simple schema of inclusion and exclusion criteria. This schema includes options like private room, private house, how many rooms available, how many bathrooms available, and host ratings and reviews among many other options.

Through the use of GIS data mining, we will provide an innovative yet simple solution centered on patients to find the most suitable trail in the closest proximity to them. Our system will not only allow more control of the clinical trial by the patient and the researcher but it will also save money and some of the hardships experienced by the patients. This will certainly lower the threshold for participation to different trials.

The use case of GIS data mining will be as follows (Figure 2):

Our software model will extract required information from different clinical trials including databases like https://clinicaltrials.gov/. Retrieved data including information

about active recruitment centers for a particular trial as well as addresses for each center will be stored in the database. The extracted data will be created and stored using HL7 FHIR standards by our data assembly and machine learning teams. This database will always be available in the investigator interface to be updated on an as needed basis. Our quality assurance team will also have access to update the database of which centers or hospital are actively recruiting.

Once the patient or his/her medical care provider enters the patient information the software will then extract the following key information: type of clinical trials patient is seeking, medical diagnosis and patient's address. It is safe to say that not all providers are using an EMR system that supports FHIR standards and we are expecting that this will be one of the major challenges for our application. This is why we are mitigating this challenge by the use of the NLP layer of the application. If the provider EMR does not support FHIR standards, scanning and sending the medical records in images of PDF files will be the other data portal for our application. After which the rest of the tires (neural network and GIS data system) will work again the same fashion to provide the best available matching. The software will then query all actively recruiting centers database to match the patient with the clinical trial in the closest geographical proximity based on two constrains: type of clinical trials selected by the patients and patient's medical diagnosis. It is worth noting that our GIS data mining algorithm will follow a relatively simple GIS data mining technique using mainly the Zip code for the propose of simplifying the application. We do not need a very robust algorithm that creates instant ongoing GIS tracking queries like those used in the UBER platform which allows

continued detection and monitoring between the drivers and customers in much more detail more than just Zip code.

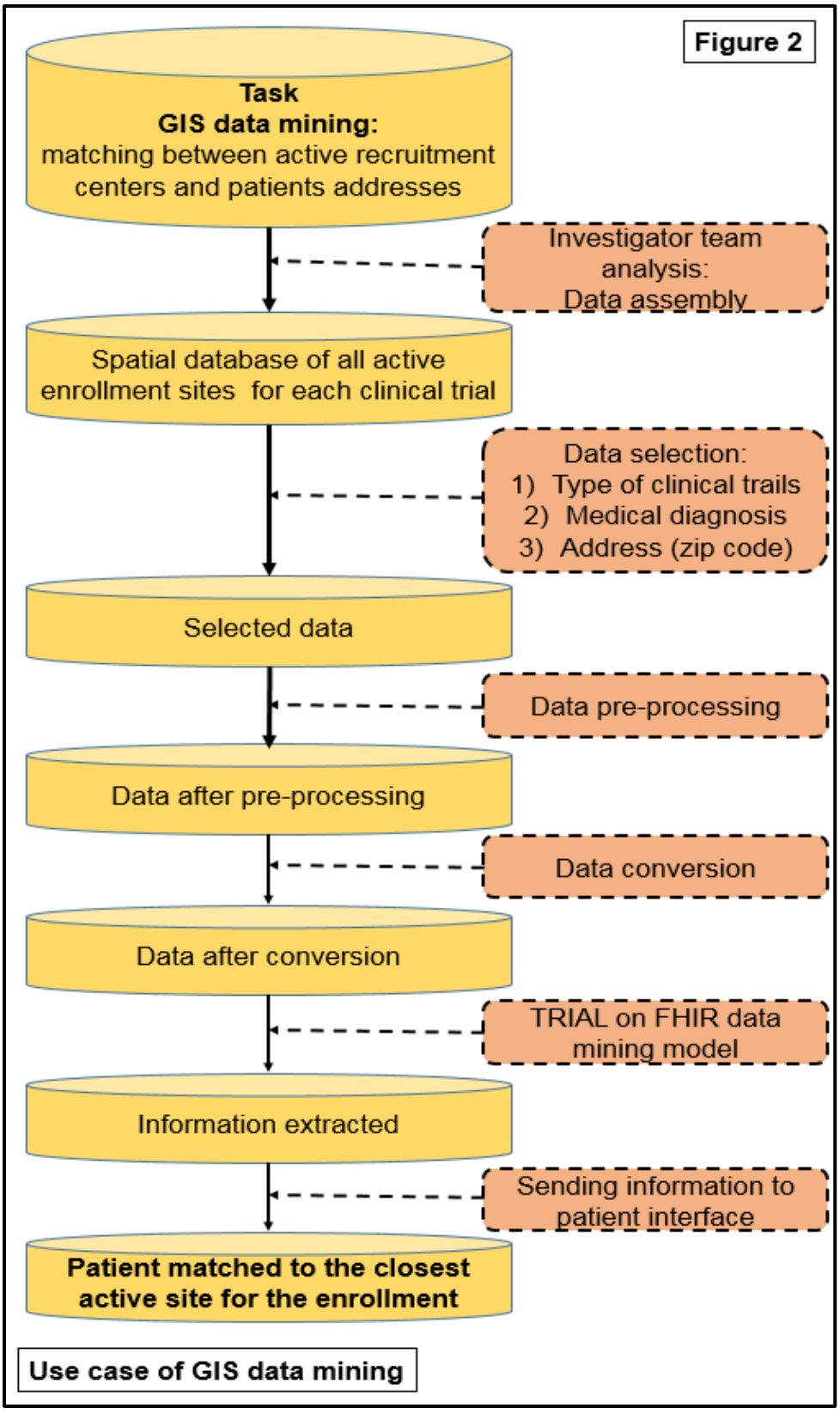


Figure 2
Task
GIS data mining:
matching between active recruitment centers and patients addresses
Investigator team analysis:
Data assembly
Spatial database of all active enrollment sites for each clinical trial
Data selection:
1) Type of clinical trails
2) Medical diagnosis
3) Address (zip code)
Selected data
Data pre-processing
Data after pre-processing
Data conversion
Data after conversion
TRIAL on FHIR data mining model
Information extracted
Sending information to patient interface
Patient matched to the closest active site for the enrollment
Use case of GIS data mining

3.1.2. Machine learning

Machine learning is the process of learning patterns from available data to make predictions that generalize to "future unseen" data. It is generally divided into two major types: supervised and unsupervised learning. When labels are available for the dataset, a supervised learning approach is often used to learn how to predict these labels from the features provided. When labels are not available, an "unsupervised" approach is used, where there is no phenotype or outcome to predict, but a supposed underlying structure of the data is being discovered.

A fundamental concept in machine learning is data separation and the quest for generalization. Before making any predictions, the data is divided into training and testing sets. The testing set, also known as the "held out" set, is used to test how generalizable the trained model would be if it were to be used on future unseen data. A simple example to illustrate this is polynomial fitting. Suppose we have two synthetically-generated random variables (X1 and X2), which when drawn in a scatter plot (with X1 and X2 being the two axes) have no underlying pattern. Given a polynomial fitting algorithm, it is possible to explain much of the variance of the data with a very high-degree polynomial, given enough training iterations. In other words, without restraint on model complexity, it is possible to explain almost any dataset to an arbitrary level of accuracy. This does not mean, of course, that the model will have any meaning or generalization, and indeed our high degree polynomial is very unlikely to be even close to accurate when it is applied to the testing data. This is known as model *"overfitting"*. There is a well-known trade-off between model fitting and generalization, and there almost always exists a "sweet spot" where the model fits the training data well

enough to have any meaning, but is generalizable enough to allow for utility over future unseen data.

Most of the machines learning algorithms require tuning of model "hyperparameters." In regularized linear models, for example, it is necessary to determine how much to penalize the weights, and in neural networks, it is necessary to determine what network architecture and learning behavior, including the number of nodes per layer (width), the number of layers (depth), the learning rate, the type of non-linearity and the type of optimizer to use. If we were to tune these parameters on the testing set, we would be defeating the purpose of an independent, untouched testing set to evaluate model generalization. Therefore, the training set is often further divided into training and validation sets, where the validation set is used to provide feedback on model generalizability to choose the most optimum hyperparameter configuration.

We are proposing the use of a Neural Networks machine learning approach to train the matching algorithm. This choice is based on recent successes neural networks have had in handing some of the most difficult prediction tasks, especially in the image classification, natural language processing and time series fields. Machine learning using neural networks with more than one layer is often called *Deep Learning*, and has exploded in popularity over the last century because of the advent of novel analytic/computational approaches as well as the increased power of computing and availability of "Big Data." Originally modeled to mimic the human Central Nervous System (CNS) and later taking on a life of its own, neural networks are fundamentally very simple. Each node represents the result of a linear combination of input features, with different weights and added biases, followed by the application of a non-linearity

(known as the *activation function)*. The most common nonlinearities used are sigmoidal, Tanh and Rectified Linear Unit (ReLU). Logistic regression is, essentially, a single node with a sigmoidal activation function (also known as a perceptron). The power of neural networks comes from combining multiple perceptrons in clever ways horizontally (forming nodes of a single layer) and vertically (forming multiple layers) to approximate functions to an arbitrary level of accuracy. The key algorithm that resulted in the explosive rise of deep learning and that made neural-based learning mainstream is *backpropagation*, the process of using the multivariate chain-rule to update all parameters (i.e. all weights and biases) simultaneously after each pass through the network, thus saving considerable time and effort (18,19).

Multiple reasons exist for our preference of using neural networks (with NLP) in this project. Our algorithm will superimpose neural network modeling on textural NLP. The approach provides two strong artificial intelligence tools that synergize each other. The model is expected to extract textual data from www.Clinicaltrials.gov, which given the serial nature of text is particularly well-suited to Recurrent Neural Networks, specifically the LSTM (Long-Short-Term-Memory) variant. Not only can neural networks capture the sequential dependency between textual elements, they have the added advantage of being *representation/feature learner*(20)*s.* "Classical" machine learning involves initial feature extraction and feature engineering, a process which is highly dependent on individual expertise and is not guaranteed to be optimum or reproducible. On the other hand, neural networks are ideal in situations where it is not very clear what features should be extracted, since the backpropagation algorithm will

modify the weights and biases in such a way to automatically capture and emphasize relevant components of the feature space and to de-emphasize everything that is not.

The most common criticisms of deep learning is that neural networks are "data-hungry," prone to overfitting and act as a "black-box" algorithm with difficult model interpretation. These drawbacks should not present an issue in this project given:

1) The abundance of clinical trial data
2) The advent of innovative regularization and sparsity-encouraging techniques including *dropout* and the *ReLU* activation function
3) The emphasis of this project on model accuracy (correctly matching patients to the relevant trials) rather than model interpretability(21).

However, we have to add that some nodes of the proposed neural network will use a 'simple' and supervised approach for different queries. A simple 'decision-tree' node where there are strict trial exclusion criteria (e.g. younger than 18 years old = no) may be used to simplify our model. This is one of the great features of using the neural network as it acquires a high threshold of flexibility that allows creating different nodes with different functionalities.

3.1.3. NLP data mining

Text data mining or natural language processing is the process of extracting high-quality meaningful information from the available text in different document format files. This process involves the interface between computer languages and human (nature) languages in order to produce meaningful knowledge to a certain required query (22,23). By utilizing NLP (with neural network), we can implement relationship extraction and sentiment analysis to match patients and clinical trials through extracting phrases from

patient's medical records (patient side) and inclusion and exclusion criteria (investigator side) and create a matching value between both data sources. NLP works by two main components, Natural Language understanding (NLU) and National Language Generation (NLG). NLU is the process of providing a meaningful value (output) of the natural (human) language that the algorithm received. This happens by breaking the available texts into small units, then compare these units to pre-trained units in the memory in order to learn what these small units mean and what is the relationships between these units and each other. The second step, the NLG process, involves the translation of the output of NLU into another natural (human) text that is meaningful to the original required query(24).

There are many tasks that NLP can be used for. Here we are only focusing and referring to a certain task of NLP which is Deep analytics**.** Deep analytics is the task of data processing from multiple source documents to allow the extracting of valid specific knowledge. Deep analytics is important when targeting highly comprehensive queries with glandular source documents that are both structured and unstructured. It is used in different sectors including social network feeding, trends, Rich Site Summary (RSS), biomedical industries and data mining for certain business values(25).

As we mentioned earlier, our algorithm will superimpose neural network modeling on textural NLP. The approach provides two strong artificial intelligence tools that synergize each other. This approach had gotten more popular recently with very encouraging outcomes (26).

Neural Networks are very robust learning approaches that can adapt to different learning model from supervised to unsupervised and from liner to non-linear models. On the other

hand, NLP algorithms are created mainly for machine learning models. They rely on analyzing a set of examples (collection of highlighted sentences), then making statistical inferences by analyzing these examples. They are mainly used for summarizing script blocks, creating chat bots, creating keyword tags, identifying entity types, identifying strings, generating words roots and creating text tokens(23).

NLP can deals with different sizes of structured and non-structured data. Our model will work by capturing the similarities between different documents and propose suggested outcomes from these similarities. Each outcome from NLP will transcend to another layer in the neural network model to process another layer of functionality. Various models of neural network can be used for drop-in replacement fashion to predict different synonyms for different medical terms by extracting these terms through NLP. They usually work for prediction problems in binary classification, multi-class classification or complex structured prediction problems. They usually do that with high accuracy due to their capability of simultaneous and simple integration of word embedding through the previous training to the algorithm. Neural network models also showed superiority in text categorization, classification of document type, phrases identifications and recognize relationships between entities (26).

As we mentioned earlier, it is important to emphasize that not all providers with have acquired HL7 FHIR standards in their EMR systems. This can be very challenging for our software if we are relying only on systems support FHIR. This is why our model is relaying on the cutting edge technology of the NLP layer as well as neural network layers to manage this problem of the application. If the provider EMR does not support FHIR standards, scanning and sending the medical records in images of PDF files will be

the other options for data analysis. NLP will scan the PDF copy of medical records and will start the process of matching with neural network to the best suitable clinical trials. **The use case of NLP data mining with neural network algorithm will be as follows (Figure 3):**

Our software will extract different required information from different clinical trials (including https://clinicaltrials.gov/ and different pharmaceutical companies vendors listing their trials inclusion and exclusion criteria) and various patient medical records files (from different EHR systems that our model will be implemented through FHIR standards). The model will assemble all the reportable text from different resources. It will extract these report texts for each trial. The algorithm will do a sentence searching process to identify the target sentences. These sentences will be transferred into phrases-chunks, then word candidate chunks and finally candidate values for each vector. Using neural network machine learning model, our data assembly and machine learning team will teach the model the different values for each vector (cell) in the neural network model. For example, the model will learn that the word "cardiomyopathy" has the same meaning as "heart failure," and if the clinical trial inclusion criteria includes "cardiomyopathy" in it and the patient's medical records includes "heart failure," then one relationship of enrollment has been checked. Each trial will have a "candidate values" list for the targeted pre-trained meanings and relationships on the database. This database will be created by our data assembly and machine learning teams and will be created and stored using HL7 FHIR standards.

This database can only be updated by the machine learning team and quality assurance team as new updates in patient's records or clinical trials enrollment criteria

become available. Once the patient (or medical care provider) enter the patients' medical records, the software will extract these meanings and relationships in the same approach (Report text → sentence searching → Target sentences → Phrase chunks → Chunk Filtering → candidate values) The software will then query both report texts (enrollment criteria and patient's medical records) to match this patient with suitable clinical trial then with the geographically closest active recruitment center based on NLP/neural network component and GIS data mining component).

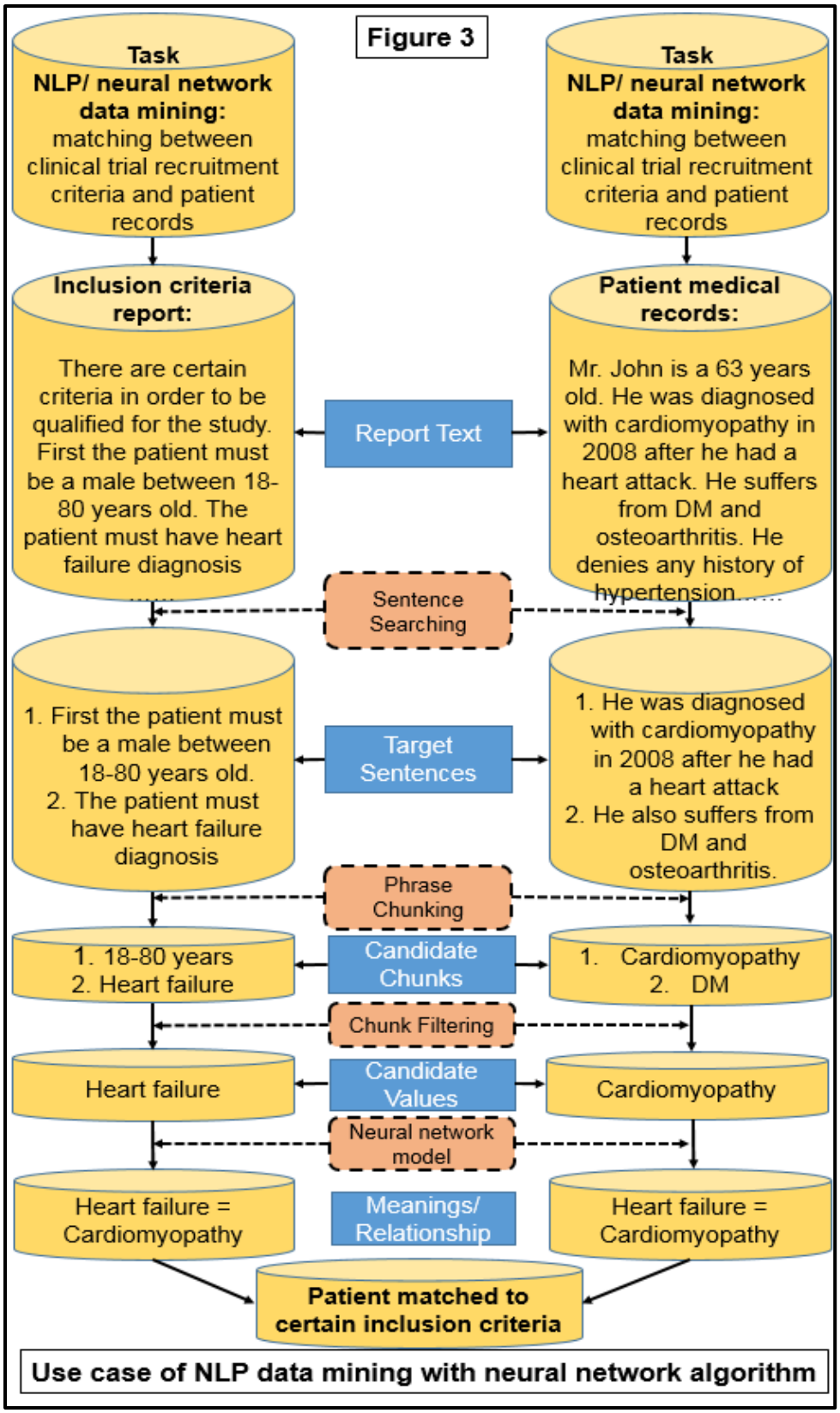


Figure 3
Task
NLP/ neural network data mining:
matching between clinical trial recruitment criteria and patient records
Task
NLP/ neural network data mining:
matching between clinical trial recruitment criteria and patient records
Inclusion criteria report:
There are certain criteria in order to be qualified for the study. First the patient must be a male between 18-80 years old. The patient must have heart failure diagnosis .....
Report Text
Patient medical records:
Mr. John is a 63 years old. He was diagnosed with cardiomyopathy in 2008 after he had a heart attack. He suffers from DM and osteoarthritis. He denies any history of hypertension.....
Sentence Searching
1. First the patient must be a male between 18-80 years old.
2. The patient must have heart failure diagnosis
Target Sentences
1. He was diagnosed with cardiomyopathy in 2008 after he had a heart attack
2. He also suffers from DM and osteoarthritis.
Phrase Chunking
1. 18-80 years
2. Heart failure
Candidate Chunks
1. Cardiomyopathy
2. DM
Chunk Filtering
Heart failure
Candidate Values
Cardiomyopathy
Neural network model
Heart failure = Cardiomyopathy
Meanings/ Relationship
Heart failure = Cardiomyopathy
Patient matched to certain inclusion criteria
Use case of NLP data mining with neural network algorithm

3.1.4. HL7 FHIR standards

As mentioned earlier, FHIR is both standard resources for data formatting and API that can be used to define methods of communications for the exchange of records between different EHRs. It supports wide a spectrum of protocols and data formats including HTTP, HTML, JSON and XML among others.

FHIR works by revealing the targeted data elements directly as API services. This is why it can develop different user-friendly applications. It will allow the development of an API model with two interfaces (patient and investigator) by simultaneously allowing access of eligibility criteria from the investigator side and extraction of pertinent information in medical records from the patients' side in automated fashion.

FHIR works by dividing each element of healthcare records, such as patient's demographic data, admission, diagnostic reports and medications, and transcribing them into their own messages that can be transferred using different data formats (JSON, XML or TXT). This feature of classification, deployment and then aggregation of FHIR is the golden advantage that we will use for our model. These newly created small data formats will be targeted and matched between the medical records from the patient and trial information from the investigator/researcher.

The use case of HL7 FHIR standards implementations will be as follows:

The first step will be the deployment of medical records and clinical trial information into small FHIR standards using one of the data formats (most probably XML for its

flexibility and compatibility with different HTTP and HTML protocols). After deploying different source documents into FHIR standards, the model will then assign each FHIR standards from the patient side to the corresponding one from the investigator side. For instance, the model will assign the patient address a FHIR standards and assign the investigator FHIR message for location of active sites. It will also assign patient demographic information a FHIR message (including age, sex) and assign the investigator a FHIR message including age and gender criteria for enrollment. Once the model stratifies and assigns all these HL7 messages to their corresponding message, components of our data mining model (GIS, NLP/neural network) will start matching the patient to the clinical trial, and finally the model will create a probability percentage of a possible match. We will use cutoff for match percentages as follows:

1) If the match is 70-100% compatible then the model will display the one-on-one result to the user interface. It will send the information to the patient (or his/her provider) regarding the clinical trial including the closest active site, the contact information of clinical coordinator responsible for the trial and the benefits (both medical and not medical incentives).
2) If the match is 50-70% then the model will display the top three compatible results (best suited clinical trials) based on the inclusion and exclusion criteria. It will show a list of required information that is missing between, or possible mismatched information between the patient records and the recruitment criteria. It will also show the closest active site, the contact information of the clinical coordinator responsible for the trial and the benefits (both medical and not medical incentives) for each trial.

3) If the match is less than 50% then the model will display non-availability message. However, patient's medical records will stored in the database and a query can be made again if the database is updated in hopes of getting a match, only after the patient gives a permission for the rematch process.

Additionally, our model will deal with highly sensitive and private individual medical information that is protected under HIPAA regulations. In order to insure the compliance of our model to HIPAA regulations, our model will process database transactions that transfer patient medical records in an ACID fashion (Atomicity, Consistency, Isolation and Durability). This means that it can only happen after an approved consent has be obtained through the app from the patient user tier.

. Finally, the database transaction from the investigator interface side will occur in a BASE fashion (Basic Availability, Soft-state, Eventual consistency).

To illustrate the use HL7 FHIR standards we are presenting a sample of HL7 FHIR data elements. We will illustrate a matching process between the data elements of common clinical data set (CCDS) along with the enrollment criteria for a clinical trial. First we came up with enrollment criteria for a clinical trial. Our clinical trial will aim to study the efficacy and safety of a new pharmaceutical medication (Agent X) in treatment of chronic heart failure patients in males. Below is the enrollment criterion for our trial:

- Inclusion criteria:
    1) The subject must be 18 to 80 years of age at screening
    2) Male and female patients can participate
    3) Any race or ethnicity can be enrolled in the study
    4) Must be able to understand and speak English language fluently

5) Diagnosis of the ischemic or non-ischemic heart failure at least 6 months prior to screening
6) Patient is on optimum stable doses of heart failure medications for the last 60 days before screening
7) Patient has BNP enzyme level less than 100 pg/ml
8) Patient had at least one left heart catheterization procedure in the last 2 years.
9) Patient received flu vaccination last year
10) Automatic implantable cardioverter-defibrillator (AICD) must be implanted in the patient at least 6 months prior to the screening visit
11) The patient must receive the appropriate standard of care in the last 6 months for his heart condition based on the American Heart Association (AHA) guidelines.

- Exclusion criteria:
  1) Current smoking status
  2) Patients with known current malignancy
  3) Any changes in heart failure medications in the last 60 days before screening
  4) Patients receiving immunosuppressant treatment
  5) Patient has known hypersensitivity to (substance X) which is an active component of (Agent X)
  6) Patient has hemoglobin level less than 8 g/dL
  7) Patient has a body temperature more than 38.0 C at the time of screening visit
  8) The patient is likely to be on heart transplant list in the next 6 months

The demonstrated “data element query” table below illustrates the end product matching process between the HL7 FHIR messages and the clinical trial enrollment

criteria. This matching process is based on the Argonaut Data Query Implementation Guide Version 1.0.0 (27)

No	CCDS Data Element	FHIR Resource		Enrollment Criteria
(1)	Patient Name	Patient		NA
(2)	Sex	Patient		Inclusion no. (2)
(3)	Date of birth	Patient		Inclusion no. (1)
(4)	Race	Patient		Inclusion no. (3)
(5)	Ethnicity	Patient		Inclusion no. (3)
(6)	Preferred language	Patient		Inclusion no. (4)
(7)	Smoking status	Observation		Exclusion no. (1)
(8)	Problems	Condition		Inclusion no. (5) Exclusion no. (2)
(9)	Medications	Medication, Medication Statement, Medication Order		Inclusion no. (6) Exclusion no. (3) Exclusion no. (4)
(10)	Medication allergies	Allergy Intolerance		Exclusion no. (5)
(11)	Laboratory test(s)	Observation, Diagnostic Report		NA
(12)	Laboratory value(s)/result(s)	Observation, Diagnostic Report		Inclusion no. (7) Exclusion no. (6)
(13)	Vital signs	Observation		Exclusion no. (7)
(14)	(no longer required)	-		NA

No	CCDS Data Element	FHIR Resource		Enrollment Criteria
(15)	Procedures	Procedure	⇨	Inclusion no. (8)
(16)	Care team member(s)	Care Plan	⇨	NA
(17)	Immunizations	Immunization	⇨	Inclusion no. (9)
(18)	Unique device identifier(s) for a patient's implantable device(s)	Device	⇨	Inclusion no. (10)
(19)	Assessment and plan of treatment	Care Plan	⇨	Exclusion no. (8)
(20)	Goals	Goal	⇨	NA
(21)	Health concerns	Condition	⇨	NA

3.2. Application requirements:

- **Business requirements**

1) Trail on FHIR application must support the objective of matching candidate patients to the closest suitable clinical trials available.
2) The application must increase the recruitment rate for participated clinical trials.
3) The application may increase public knowledge regarding the benefits of the use of clinical trials.
4) The application should improve our data quality by implementing ongoing evaluation
5) The application should provide an easier interface for patients to review clinical trials.
6) The new system should provide an easier interface for researchers to find suitable candidates.

7) The application must have the flexibility to adapt changes overtime

- **Functional requirements**

➢ **Organization/Provider management**

1) Contact information (e.g., name, address, contact numbers, unique identifiers) of all available clinical trials sites must be captured and stored in the database.
2) The application must include the location of the clinical trial.
3) The application must support data elements

➢ **Patient management**

1) The application must capture the patient demographic information including first name, last name, date of birth and gender.
2) The application must capture patient current address
3) The application must capture patients' medical records in any types of file formats.
4) The application must assign probability percentage between patients and suitable clinical trials.

➢ **Clinical trial management**

1) The application must identify clinical trials' inclusion criteria
2) The application must identify clinical trials' exclusion criteria
3) The application must identify clinical trials' feasibility
4) The application must identify clinical trials' side effects
5) The application must store information about clinical trial sponsor company
6) The application may track clinical trials' related scientific publications

➢ **Administration and system interface management**

1) The application must provide patients specific medical information to corresponding research investigator, those with patients consent.
2) The application must send courtesy emails to patients once a suitable clinical trial becomes available for their condition
3) The application must generate an official enrollment status report for each research investigator

- **Non-functional requirements**

➢ **Usability requirements**

1) Both APIs must be user-friendly.
2) The application platform must be available for users 24/7 a week.
3) The application must allow research investigators to create user accounts to enter their trial information
4) The system must allow patients to create user accounts to enter their personal and medical information
5) The system must be able to work on a PC or tablet, supporting multiple common operating systems for mobile devices

➢ **Performance requirements**

1) The application database must provide sufficient storage for current information
2) Database must be scalable to accommodate huge influxes of new data.
3) The application must support data electronic exchange of records using HL7 standards for message format and contents
4) The application must have a strong and high performance operating system.

➢ **Supportability requirements**

1) The application must include on-line help service.
2) The application must support direct communication between users through emails.
3) The application must support a tracking structure for fixing potential bugs in different IT tiers.

- **Security requirements**

1) The application may require two-factor authentication.
2) The application must acquire strong firewall to prevent cyber threats.
3) The application must require strong guidelines for creating passwords for users including letters, numbers, capital letters and special characters.
4) The application must require changing of the user password every 6 months.

- **Transition requirements**

1) The application must support the smooth transition between different stages in the project lifecycle

- **Compliance requirements**

1) The Application must follow the HIPAA regulation to patients' sensitive information.
2) The Application must comply with the Collaborative Institutional Training Initiative (CITI) training and guidelines
3) The Application must comply with the Key concepts and Good Clinical Practice (GCP) guidelines
4) The Application must comply with the Institutional Review Committee (IRB) guidelines
5) The new system must comply with different programming language rubrics, including R language and JavaScript

6) The new system must be in compliance with security requirements for Certification and Accreditation (C&A)

3.3. Data elements:

The two main data

entities are patients' medical records and clinical trial inclusion/exclusion criteria. The application must support and be able to receive any type of structural data (e.g. PDF, doc, text, HL7), or non-structural data (e.g. JPG images). The application must have a domain for scanning and transforming hard copies of the documents into electronic versions.

- **Structure data extraction**

Different algorithms for structured data extraction will be used. The comprehensive approach of all these algorithms will identify the appropriate template of the required data from relevant patient records (i.e. cardiac patients from the cardiology department for cardiology trials, neurology patients for neurology trials). Once the basic requirements are met the algorithm will then use stratifying concepts techniques such as equivalence classes and differentiating roles to identify the required data for the requested query. This approach extremely valuable, and beneficial results in extracting data from web pages (28).

- **Unstructured data extraction**

The proposed platform will work on extracting data from unstructured files such as PDF and images and combine them with a broad variety of training data (both supervised and unsupervised) from other sources. These sources include different scanned files such

as pdf or jpg. We will use a similar approach to that being by IBM Watson to tackle different diseases through medical imaging (29).

3.4. Data architecture:

The application addresses the analysis of "big data" through supervised and unsupervised machine learning models. In order to align, store, access, analyze and synchronize all these attributes we need to create a strong data platform that supports both relational database management system (R DBMS) as well as big non-relational data. The theoretical architecture of this platform will be in a *cylindrical* design with RDBMS at the core to handle the Relational database with Apache Hadoop surrounding it to analyze and deliver Non-Relational, NoSQL data to API/data warehouses on the surface. The idea is to create a strong table-based database for the main elements of the initiative. Collecting and processing the data from different sources will be Hadoop based.

For the IT Infrastructure backbone of the "Trial on FHIR" application development we will be using an end-to-end cloud (hosted) platform. Primarily our solution is IaaS (Infrastructure as a Service) based architecture which is a proven scalable, flexible and cost-effective solution. Our vendor of choice will be Amazon Web Services (AWS).

Software Platform

Amazon AI platform and frameworks offer Apache MXNet for scalable deep learning models which include neural networks and integrates natural language processing (NLP). MXNet will be used to define, train, and deploy systems across a

broad range of use cases. As more data becomes available, the application custom AI models will be reassessed, fine-tuned, retrained and redeployed.

The mobile app development will utilize Amazon Mobile SDK which supports iOS and Android platforms. It allows direct access to AWS services like Amazon Lambda functions, S3 (storage), DynamoDB database. Additionally, the SDK can seamlessly integrate with Amazon AI minimum custom development through plug and play capabilities.

The app implements FHIR REST services through API Gateway which that acts as a "front door" to access data, business logic, or functionality from the application's back-end services, such as processes running on Amazon Elastic Compute Cloud (EC2), code running on AWS Lambda, or EHR web integrations.

Hardware Platform

Our hardware of choice will be Linux-based Amazon Deep Learning AMIs (Amazon Machine Images). These AMIs deploy Elastic Compute Cloud (EC2) instances. To start training the AI models, we anticipate using the 64-bit AMIs of type p2.xlarge optimized for machine learning and high-performance databases. The p2.xlarge instances comprise of 1GPU, 4 vCPUs,61GB memory and 12GB of onboard GPU memory. During the lifecycle of the application, we anticipate to scale up to p2.8xlarge instance based on demand occasionally.

3.5. Implementation computer languages:

1) Programming and exploratory analyses

- The application will use R language for programming. R is an open source programming language with a software statistical package and graphics-supporting

package. This will allow for a strong analytic probability with the capacity of graphic production. Additionally, R language demonstrates a strong compatibility with both Java and Python languages (30).

- In addition to R language, we are going to use JavaScript for coding and programing. JavaScript is a comprehensive and very dynamic run-time language that is one of the core components of the World Wide Web (www.) content. JavaScript is almost universal for most websites and user interactive applications. It is very flexible and can work with other applications and programs such as R language. It has its API for different functions including texts. It also works for non web-based extensions such as PDF functions (31).

2) SQL language (Structured Query Language)

Since we are building an immense database to store, analyze, query patients' medical records and pharmaceutical trials enrollment criteria, we need a strong programming language for designing and managing RDBMS. This language is essential as it allows the extraction and/or query of different information we need from our database. SQL language consists of three main subtypes of languages: data definition language (DDL), data manipulation language (DML) and data control language (DCL). Together they illustrate the scope of SQL language in creating, querying and updating the database. They consists certain data commands such as: Select, Insert, Query, Update and Delete(32).

4. Business case description:

4.1. Purpose of project charter

Constructing a phone application with an online database to match different geographically allocated clinical trials to patient's interest and medical conditions living in a close proximity to a trial. This application will be built based on FHIR platform.

4.2. Project and product overview

Who: TRIAL ON FHIR will be a new developing phone application.

What: This application will be able to match clinical trials to suitable patient candidates.

Where: The matching process will be based on geographic location of both the university hospital and community hospitals providing the clinical trial and the address of the candidate patient.

Duration: The estimated time to launching the application is 38 months. The project is ongoing after that.

Budget: $ 5 million (estimated figure, TBD further)

4.3. Justification

TRIAL on FHIR is a new fictional project. The project aims to create a new application for smart phones and tablets. We will try to illustrate the problem that this app will tackle, the current approaches of dealing with it, the proposed approach through this application, how to implement it, cost-benefit assessment and anticipated milestones.

Different clinical trials are conducted by enrolling suitable patients with a specific medical condition based on a specific list of sponsor inclusion and exclusion criteria. Usually, a clinical research coordinator is responsible for this process and does so by going through the medical records of each potential patient available to him/her in his/her medical institute database. At the same time, many patients especially with advanced medical conditions with bad prognosis are searching for innovative treatment options that

can help with their condition and potentially prolong their lives. However data shows that there is a mismatch between patient interest and the number of patient recruited.

We are proposing the creation of a mobile application to match patients with a suitable clinical trial based on both his/her medical condition and his/her geographic location. The application functionality is similar to how the UBER application works. The patient or his designated medical care provider (e.g. primary care physician PCP) will enter the patient information on the app. The app will then match the patient with the suitable clinical trial that focuses on his/her medical condition in the closest geographic location.

4.4. Business case objectives

1. Getting stakeholders approval
2. Creating database
3. Creating the application different tiers and application coding
4. Launching application

4.5. High-level requirements

The following table presents the requirements that the project's products, service or results must meet in order for the project's objectives to be satisfied.

Req. #	Requirement Description
Getting stakeholders approval	Providing business case proposal to different stakeholders
Creating database	Receiving different database requirement from stakeholders (e.g. HL7, enrollment criterial from clinical trials)

Req. #	Requirement Description
Creating application different tiers and application coding	Creating different IT essential requirements to create these tiers and performing the coding process
Launching application	Launching the product on different software platforms (iOS, Android, PC, online database, website)

4.6. Major Deliverables:

The following table represents the major deliverables that the project's product, service or result must meet in order for the project objectives to be satisfied.

Major Deliverable	Deliverable Description
Getting stakeholders approval	Receiving agreements from different stakeholders to participate in the project
Creating database	Receiving different resources from stakeholders: HL7/FHIR, Clinical trial database, IT support, university hospital database)
Creating application different tiers and application coding	Processing of application database, application user tier logic, user interface and implement coding
Launching application	Providing application testing, training and launching the product on different software platforms

4.7. Executive milestones:

Executive Milestones	Estimated Completion Timeframe
Project planning and writing proposal	6 months to receive all approvals
Receiving FHIR required resources	6 months after receiving all approvals
Conducting start-off meeting	1 month after receiving FHIR required resources
Creating different application tiers	12 months from start-up meeting
Testing and training	12 months after creating application tiers
Launching the application	1 month after testing and training.

4.8. Budget estimate:

The funding of the project will be supported by grants from different stakeholders.

Category	**Description**	**Total Amount**
Personnel: Salaries and Professional Fees for Trail on FHIR. Coordinators	Administrator of the grant	$1,750,000
IT informatics personnel	Oversee enrollment of patients	$1,750,000
Subtotal Personnel		**$3,500,000**
Amazon EC2 Amazon Aurora Amazon Redshift Amazon EMR (Apache Hadoop) AWS Cloud Storage Tableau Server	Different database center website including 10 high processor PCs and high speed internet	$ 73, 044 $42,515 $53,800 $34,470 $252,000 $31,500

DUO Web Adobe Dreamweaver Amazon Cloud Fron		$8,100 $720 $113
Subtotal Cloud based		**$496,292**
IT infrastructure fees	**Hardware host with software development**	**$620,700**
Training	**Motivational interviewing, case management training, technology training**	**$309,500**
Advertisement		$250,000
Total		**$ 5,176,492**

4.9. Assumptions:

- Patients receiving treatment in community hospital are interested in new approaches of treatment through clinical trials.
- Patients will be able to internet accessible PCs, phones, or tablets to apply for the project
- Clinical trials companies need new venues for collecting gathering participants in their trials.

4.10. Constraints:

- Receiving approval from different stakeholders
- Applying different resources for creating the database
- Advertisement for the project through different media
- Conducting the research based on HIPPA regulation

4.11. Scope borders:

- The application will inclusively present the closest geographically available clinical trials for suitable candidates
- The application will not provide any medical advice and only provide HIPAA protected information once the patient has signed the consent.

4.12. Project organization

- **Roles and Responsibilities:**

This table describes the main key roles for supporting the project.

Name	**Project Role**	**Project Responsibilities**
Initiating team	Proposal editor	• Conducting and scripting a full detailed proposal for the project
Initiating team	Stakeholder finder	• Searching for different stakeholders that can participate in the project
Initiating team	Proposal negotiator	• Negotiate different requirement from stakeholders
Initiating team	Administrative assistant	• Providing different logistic support for administrative purposes
Daily activity performing team	Project manager	• Person who performs the day-to-day management of the

Name	Project Role	Project Responsibilities
		application developing and has specific accountability for managing the project
Business team	Business analyst	• Person in response of the business flow of the project.
Data assembly and machine learning team	IT developer	• Person in management of database application, • Develop the machine learning model including user interfaces, GIS data mining, NLP data mining, neural network deep learning and coding development.
Initiating team	IT Security administrator	• Special IT developer for creation and maintaining of different security layers for the application.
Data quality assurance team	IT developer	• Assure the update of the database from both the investigator side (the enrollment criteria and active enrollment sites) and from the

Name	Project Role	Project Responsibilities
		patient side (updating medical records)

4.13. Stakeholders (Internal and external)

1) The first identified stakeholders are patients and patients care providers who are interested in participating in clinical trials.
2) EMR industry partners: these include, but not limited to the following: Accenture, Athenahealth, Beth Israel Deaconess Medical Center, Cerner, Epic , Intermountain Healthcare, Mayo Clinic, MEDITECH, McKesson, Partners HealthCare System , SMART at the Boston Children's Hospital Computational Health Informatics Program and The Advisory Board Company
3) HL7: as we are proposing using Fast Healthcare Interoperability Resource (FHIR) platform, Health Level seven (HL7) organizations will be a major stakeholder as it is the designer of FHIR.
4) U.S National Library of Medicine (NLM) at National Institute of Health (NIH): as they are the administration for clinicaltrail.gov, largest database available for clinical trials.
5) Major pharmaceutical companies that conduct human clinical trials will be reached as possible stakeholders
6) Different Informatics institutes that work by providing and conducting public health informatics to different entities.
7) Other interested stakeholders will be identified through proposal to different grants that will share interests with this initiative.

5. TRIAL on FHIR Phases (Project lifecycle)

5.1. Initiating phase (12 months):

The initiation process includes the following:

1) Writing project proposal: writing an executive application, explaining the application need, functions, methods and outcomes.
2) Identifying stakeholders: including patients, care providers, research investigators, informatics institutes, and pharmaceutical companies
3) Identifying project sponsor: the project funding needs will be identified through proposal to different grants that will be interested in this initiative. The previous mentioned stakeholders are suitable candidates as target funding sources. Different proposals will be written for the previous mentioned health institute. Finally, the special financial interest will be non-profit organization.
4) Development of business case: including background, justifications, current problems, assumptions, objectives, hypothesis, deliverables, and scope and project organization.
5) Establishing the project requirements: (see before)
6) Establishment of project expectations: evaluating the delivered service of the application based on:

- GIS: extent of covering geographical areas by the application.
- Medical specialties: extent of medical specialties covered by the application (e.g. oncology trials, cardiovascular trials, pediatric trails)
- Quality of services: ability to extract/abstract data using HL7 standard language from different clinical trials and matching them with data provided from patient/patient PCP

7) Project charter: The charter will be written based on the affiliated partners and stakeholders. It will also be written based on the approved grant proposal. The charter will illustrate budget, project information, project manager, project's key players and the different guidelines for initiating this app. However, a major portion of the charter will be focusing on Health Insurance Portability & Accountability Act (HIPAA) regulation.

5.2. Planning phase (12 months):

1) Develop project management plan

The integration management plan should include data gathering, IT training, medical science training, policy and HIPAA training, allocation of resources, renting firm headquarters, creating the essential application Platform. The plan should also explain different rules for different employees and job descriptions and illustration of periodic progress.

2) Execute the identified requirements: (see below)

3) Define scope:

- Project objectives to the from both patients and pharmaceutical companies from the other side)
- Project tools: creating mobile application, PC software and website database.
- The tool will be supported by different templates for useful links and articles, search features and user profile creation.

- Project endpoint and deliverables goal: providing patients with the available active enrolling clinical trials sites that are in close geographical proximity to their address based on their zip code and medical history.
- Estimate resources: including funding, collaborations, database materials and human resources.
- Develop schedule and estimate durations: (see above)
- Estimate costs: (see above)

5.3. Execution stage (12 months):

1) Execution eligibility checklist completion: finishing checklist to launch the project
2) Direct project execution: Downloading and providing the app on different application stores (Apple IOs and Android provider Application Stores)
3) Acquire and develop project team:

The team includes but is not limited to: clinical research coordinator, IT informaticians, IT analyst, project manager and medical directors.

4) Assignment of team tasks:

- Clinical coordinators; oversees matching of patients to clinical trials based on HL7 message, transcription case management, tracks data collection and educational activities, responsible for program evaluation and dissemination of program outcomes
- IT informatics manger: oversee enrollment of patients into the clinical trials, tracks data collection

5) Manage stakeholders expectations:

Provide up to date data to patients about the availability of the specific clinical trial (e.g. is it closed or open for enrollment), provide a statistical analysis for pharmaceutical companies about the availability and characteristics of patients' pool (without sharing individual medical information for any patient in order to comply with HIPAA guidelines)

5.4. Project monitoring and controlling stage (Ongoing):

1) Periodically scheduled meetings: to measure progress
2) Queries assessment: fixing different ongoing queries and threatening bugs (e.g. fixing a new identified security threat for breach)
3) Control project work: providing project management plan updates and verify new scope based on the new materials provided
4) Quality control: provide new features (e.g. send medical education emails to different patients regarding their medical condition. Examples include: send reducing salt food diets to heart patients and low calorie diet plans to obese patients)
5) Report performance: these reports will be announced during periodical meetings. Additionally, they are important for potential risk assessment.

5.5. Project closing stage (6 months):

1) Stakeholder/customer feedback continuous process during the monitor phase by requesting feedback based on a score system from both patients and clinical trial sponsor companies such pharmaceutical companies. Score system could be a 5 star productivity based system to provide review of different aspects of the app.
2) Lesson learned: Pros and Cons: these will help with identifying new ideas for the app.
3) Summary of project results
4) Checklist verification: in order to confirm if goals have been meet

6. Discussion:

There exists a huge gap between patient awareness about clinical trials and clinical trial participation. The current way of conducting different clinical trials encompasses the search for suitable patients with a specific medical condition and matching them based on inclusion and exclusion criteria. Usually, a clinical research coordinator is responsible for this process. He or she does so by searching the database of medical records in his/her medical institute. Patients, on the other hand, especially those with advanced medical conditions with a bad prognosis, are continuously searching for innovative treatments that can help with their condition. Yet, with patients actively looking for new clinical trials and research coordinators trying to recruit patients, there is a mismatch between patient interest and the rate of enrollment.

In this illustrated report, we are proposing the development of a novel phone application, database, and web-based software that will allow the majority of patients—irrelevant of where they receive their medical care access—to the most advanced cutting edge experimental and investigational treatment options. At the same time, we hope our software will find solutions for medical researchers and pharmaceutical companies plagued by reduced enrollment numbers in clinical trials, and will hopefully prevent the waste of millions of dollars spent on patient recruitment, respectively.

As the title suggests, we are aiming to build a bridge to close the gap between current patients and future medicine. However, we do expect certain barriers in the development of this application. We will try to illustrate the major obstacles that may interfere with the success of this application.

1) <u>Interaction between different EMR vendors:</u>

Based on medical economics statistics, there are more than 100 EHR vendors in the United States. Each one of them has its own unique coding system, set of operational standards, and different resources and APIs. While the Argonaut project helped the initiation of uniform standardization between different EHR systems, the majority of EMR vendors have not yet attempted to do so. We realize that we will engage with different hospitals whose EMR systems don't support FHIR.

We are expecting to mitigate this challenge by both the NLP layer, as well as the neural network layer. If the provider EMR does not support FHIR standards, we will acquire medical records through scanning the medical records as PDF files then processing the data (assembly, selection, conversion) through the NLP layer.

2) IT Interoperability problems:

Conducting a public health informatics project with a multidisciplinary team with a lot of specialties carries the risk of misinterpretation between team members. Additionally, community hospitals are the primary focus of our project. There is no guarantee that these hospitals have the level of sophisticated IT infrastructure and/or have adopted the EHR system that is required as part of the Argonaut project and is needed to be able to consume or send FHIR resources messages. This is why our model is designed independently of the level of the IT sophistication of these hospitals. Our model implements the HL7 FHIR tier in its own algorithm to receive, consume, and deliver HL7 messages irrelevant of the level of IT sophistication in the source. Our model works on a glandular/individual level. It will extract the information in both structured/unstructured data elements, then will transfer it to the HL7 FHIR tier that will create the message and send it to the machine learning model

to be processed. Our model supports both interoperability as well as intra-operability between its own components: that is part of the key feature in the design.

3) Sharing:

Vendors' willingness to participate, especially those who have similar/competing products (e.g. researchmatch.org, IBM Watson) could be a difficulty we might face . We are aiming to produce our model as an open source for the public to enhance the sharing of IT knowledge.

4) Not all physicians are scientists:

Even though physicians usually keep track of the most up-to-date knowledge regarding the standard of care, and some do refer their patients to clinical trials, there are many who are not are interested in engaging their patients in clinical trials. To make matters worse, the majority of patients are not familiar with either the medical terms or the technical terms of clinical trials. This is why we are emphasizing using API, as it is a user-friendly interface for both patients and investigators, in our conceptual design in order to overcome any utilization obstacles.

5) Intellectual property barrier:

Healthcare is a business enterprise, with many vendors and stakeholders in robust competition with their peers. Therefore, the idea of data sharing, specially in the pharmaceutical industry, is not often welcomed. Pharmaceutical companies are hesitant to share their protocol regulation and enrollment criteria with a third party.

6) Inability to choose the right timing in executing.

This is why we are aiming to work with both a project manager and entrepreneur to detect the appropriate timing for each stage of the project lifecycle.

7) Data mining and machine learning problems:

Creating an appropriate algorithm to deal with immense data information is challenging. Over-fitting, mis-predication and non-labeling can cause technical problems that we need to address through a strong validation system as they happen.

8) HIPAA:

The Health Insurance Portability and Accountability Act (HIPAA) was established in 1996 to protect patient rights and secure individual health information. Our database will contain a lot of identifiable health information, and any misuse or unauthorized access might be considered a breach of HIPPA guidelines. To prevent HIPPA violation, our database transactions from the patient interface to transfer patient medical records will happen in an ACID fashion, only after obtaining an approved consent through the API from the patient user tier, while database transaction from the investigator interface side will occur in a BASE fashion.

9) Sunken cost:

Developing any project or application using FHIR carries a lot of sunken costs. The reason for this is (the substantial gaps) between developing the application, launching the application, and receiving promising results from the application. The gap between these three phases consumes three major sunk costs: time, money and human training. However, we can reduce this loss by managing the relatively short life-cycles of the application phases. These phases will be linked together in an incremental way, so that end-stage users will be involved continuously in the project until the final stage is complete. The essential need of the FHIR platform is on the supply/demand cycle. FHIR provides bigger and clearer information to both end users

(researchers and patients) in order to meet their needs. FHIR enhances interoperability between different stakeholders by providing an easier and faster way to share secure health messages.

10) Future Challenges:

Challenges center on Title 21 CFR Part 11. Title 21 CFR Part 11 is part of the code of federal regulation established by the FDA on electronic records and electronic signatures. It defines the regulations through which her, as well as electronic signatures, are considered trustworthy and equivalent to paper records (33,34). It particularly targets pharmaceutical industries, contract research organizations, and biotechnologies companies. It requires the implementation of auditing and validation on any software or IT system that deals with EHR (35-37).

This regulation in particular can be challenging because of the costs associated and the possible impracticality due to the FDA's statements regarding (enforcement discretion) in applying this regulation. Additionally, this regulation is currently under revision. These different factors have led to confusion on how and what exactly is required (37).

We will comply with the regulations once the final revision is assigned. We will keep "hard copies" of EHR and use it as the authoritative document for regulation purposes.

Risk Log:

The below table shows the major identified risks, impact of the risks on the project and Project Team response.

Risk	Description	Impact	Response
1. Unproven, Cutting-Edge Technology	Machine Learning (ML), NLP and Artificial Intelligence (AI) are inherently evolving field of Information Technology. Creating an appropriate algorithm to process with immense unstructured data is challenging.	High	We will mitigate the risk. These technical challenges will be addressed and documented throughout the course of the implementation.
2.HIPAA Security	The Trial by FHIR app will be accessing Protected Health Information (PHI) from Medical records. The app will come under the HIPAA Security Act of 1996.	Low	We will mitigate this risk. The app will encrypt the data in motion and at rest using 256-bit encryption method and VA Authentication Federation Infrastructure (VAAFI) single sign

			on (SSO) service for sign-up and sign-in authentication to the mobile and web apps. Additionally, the security incident event management (SIEM) will be utilized for real-time access and monitoring unusual activities like cyber-attacks, advanced threats & malware.
3.Time/Cost Overrun	Trial on FHIR app uses a number of different innovative software and will be exploring uncharted territory of integrating various technologies.	High	We will mitigate the risk. We will use "agile" project development methodologies. We aim to develop the

	Estimating the project development timelines are challenging. This could entail significant time and cost overruns.		project in modular fashion with the aim to create incremental delivery of working software in close collaboration with partner health care organizations.
4. Data sharing	Healthcare is a business enterprise, with many vendors and exclusive partnerships. Pharmaceutical companies maybe hesitant to share data with a third-party app.	Medium	We accept the risk.
5. Health partner r referrals	Physicians or hospital may not be interested in referring their patient or advising their patient to engage in clinical trials. This can be a barrier to	High	We intend to mitigate the risk. We will work with the pharmaceutical industry to advertise

	finding the patients to match to the clinical trials.		the application through direct channels so as to reach the physicians at their clinics and hospitals. We will partner with Hospital clinical trials to conduct education for physicians and providers for the Trial by FHIR app.

Risk Monitoring, Controlling, and Reporting:

The level of risk of the project will be tracked, monitored, and reported throughout the project lifecycle. A "Top 10 Risk List" will be maintained by the project team, and will be reported as a component of the project status reporting process for this project. All project change requests will be analyzed for their possible impact on the project risks. Management will be notified of important changes to risk status as a component to the Executive Project Status Report.

7. Executive summary:

- **Introduction:**

Every day scientists push the boundaries to discover new treatments. Treatments like antibiotics and vaccines bring about profound changes in the world. But unfortunately, many of these new treatments never pass the testing stage. The two main reasons for clinical trial failures are: a steady decline in the number of patients recruited and failure to maintain patients in the trial. It is estimated that recruitment rates in clinical trials have dropped 20% since 2000, largely due to the fact that many patients lack knowledge about clinical trials that may be suited for their medical conditions. On the other hand, a large number of patients receive medical care at doctors' offices or community hospitals with fewer research opportunities, limiting the patient pool available to those researchers conducting trials at university or urban hospitals. Finally, current options to enhance recruitment and retention do not take patients or clinical trial sites into consideration while matching patients with researchers.

- **The necessity:**

Patients with chronic diseases usually feel that they are fighting an uphill battle against their disease. They feel hopeless facing their disease, especially if the standard of care is not optimal for their situation. Clinical trials provide patients with options that would otherwise not be part of their standard of care, and in doing so they give patients a feeling of control over their disease and hope for the future. But problems with recruitment and retention continue to plague researchers. In 2010, an estimated fifteen billion dollars was spent to enhance patient recruitment and the retention of clinical trials. Yet the number of

patients recruited has declined over the years. There is an urgent need to bring patients and researchers together in new ways.

- **The proposal:**

We are proposing the development of a novel phone application and a web-based software interconnected through a database that will match patients with cutting edge experimental and investigational treatment options based on their medical condition and proximity to clinical trial sites.

- **Methodology:**

We plan to build a platform and software algorithm assembled on: geographic information system (GIS) data mining, neural network machine learning, Natural Language Processing (NLP) artificial intelligence, and HL7 FHIR standards to match patient's medical records with clinical trials enrollment criteria. Medical records will be provided by the patient while the pharmaceutical company or any other entity sponsoring the trial will provide the inclusion and exclusion criteria. Data in the above-mentioned source documents could be mapped to HL7 FHIR messages on an as-needed basis. Our model will apply a similar concept to UBER and Airbnb in terms of matching between users/guests and drivers/riders, respectively.

- **Conclusion:**

We plan to create a platform that we hope will provide patients access to the most current and cutting-edge experimental and investigational treatment options, and hopefully will give back some of the control that their disease has taken from them. Our software will also help to increase the patient pool available to researchers, reduce the time spent on

finding suitable candidates, and finally drive down the costs associated with conducting trials.

8. References:

1. National Institutes of Health (U.S.), National Library of Medicine (U.S.), United States. Food and Drug Administration. ClinicalTrials.gov.
2. Landray MJ, Bax JJ, Alliot L et al. Improving public health by improving clinical trial guidelines and their application. Eur Heart J 2017;38:1632-1637.
3. Johnston SC, Rootenberg JD, Katrak S, Smith WS, Elkins JS. Effect of a US National Institutes of Health programme of clinical trials on public health and costs. Lancet 2006;367:1319-27.
4. Cobb EM, Singer DC, Davis MM. Public Interest in Medical Research Participation: Differences by Volunteer Status and Study Type. Clinical and Translational Science 2014;7:145-149.
5. Detoc M, Gagneux-Brunon A, Lucht F, Botelho-Nevers E. Barriers and motivations to volunteers' participation in preventive vaccine trials: a systematic review. Expert Rev Vaccines 2017:1-11.
6. Analytics Z. National Poll on Clinical Reserach In: America R, editor National Poll: Reserach America June 2013.
7. Analytics Z. Americans' views on advancing research and innovation National Survey Methodology Research America Research America 2016.
8. Krohn T. Which Clinical Trial Matching Service is Right for You? www.lillypad.lilly.com: lillypad, 2013.
9. Cahan A, Cimino JJ. Visual assessment of the similarity between a patient and trial population: Is This Clinical Trial Applicable to My Patient? Appl Clin Inform 2016;7:477-88.
10. Tramm R, Daws K, Schadewaldt V. Clinical trial recruitment – a complex intervention? Journal of Clinical Nursing 2013;22:2436-2443.
11. Berndt ER, Cockburn IM. Price indexes for clinical trial research: a feasibility study. Mon Labor Rev 2014.
12. McGuire R. Patient Recruitment and Clinical Vendor Fees Top Clinical Trial Cost Drivers. Cutting Edge Information 2011.
13. Baquet CR, Henderson K, Commiskey P, Morrow JN. Clinical trials: the art of enrollment. Semin Oncol Nurs 2008;24:262-9.
14. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. J Clin Nurs 2010;19:227-33.
15. Kubick W. Building on FHIR for Pharmaceutical Research. Applied Clinical Trials 2016.
16. JASON. A Robust Health Data Infrastructure. AHRQ Publication 2014.
17. Kenneth E. Foote ML. Geographic Information Systems as an Integrating Technology: Context, Concepts, and Definitions. The Geographer's Craft Project 2015.
18. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. New York: Springer, 2001.
19. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction. 2nd ed. New York, NY: Springer, 2009.
20. Has¸im Sak AS, Franc¸oise Beaufays. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. Google, USA.
21. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-444.
22. Point T. AI - Natural Language Processing.

23. Kiser M. Introduction to Natural Language Processing (NLP) 2016. Algorithmia, 2016.
24. Neospeech. WHAT IS NATURAL LANGUAGE PROCESSING AND HOW DOES IT WORK? In: Neospeech, editor Text2speech blog, 2016.
25. Sims S. Everything You Need to Know about Natural Language Processing. In: KDnuggets, editor What is natural language processing? , 2015.
26. Goldberg Y. A Primer on Neural Network Models for Natural Language Processing. J Artif Intell Res 2016;57:345-420.
27. HL7 launches Argonaut Project to advance FHIR interoperability standard. Health Manag Technol 2015;36:26.
28. Arasu A, Garcia-Molina H. Extracting structured data from web pages (poster). Proc Int Conf Data 2003:698-698.
29. Fiber L. Medical Imaging Leaders Tap IBM and Watson to Tackle Cancer, Diabetes, Eye Health, Brain Disease and Heart Disease. IBM watson, 2016.
30. Hornik K. Frequently Asked Questions on R. 2016.
31. Severance C. JavaScript: Designing a Language in 10 Days. Computer 2012;45:7-8.
32. Britannica TEoE. SQL computer language. 2016.
33. Interpharm Press Inc. Code of federal regulations. Title 21, food and drugs. Interpharm master keyword guide. Reg-words series of keyword guides. Buffalo Grove, IL: Interpharm Press:v.
34. Drug USF. Part 11, Electronic Records; Electronic Signatures — Scope and Application. In: Administration F, editor, 2016.
35. TILTON CJ. The Role of Biometrics in Enterprise Security. Dell Power Solutions 2006.
36. Huber L. 21 CFR Part 11 - Electronic Records and Electronic Signatures. 2007.
37. center Mt. Food and Drug Administration CFR Title 21 Part 11. 2017.