AN AGILE RESEARCH DATA REPOSITORY OF ACUTE KIDNEY INJURY USING PROPERTY GRAPH DATABASES

By

Ahmad S. Baghal

A Dissertation Submitted to

Rutgers – School of Health Professions

in partial fulfilment of the Requirements for the Degree of

Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics

School of Health Professions Rutgers, the State University of New Jersey

January 2020

RUTGERS
School of Health Professions

FINAL DISSERTATION DEFENSE APPROVAL FORM

An Agile Research Data Repository of Acute Kidney Injury Using Property Graph Databases

BY

Ahmad Salem Baghal, MD, MS

DISSERTATION COMMITTEE:

Shankar Srinivasan PhD

Suril Gohel PhD

Corey Hayes PharmD PhD

APPROVED BY THE DISSERTATION COMMITTEE:

_____ Date: _____

_____ Date: _____

_____ Date: _____

_____ Date: _____

_____ Date: _____

TABLE OF CONTENTS

LIST OF TABLES…… vi

LIST OF FIGURES…….. vii

CHAPTER I INTRDOCUTION…… 1

A. BACKGROUND OF THE PROBLEM ……..…… 1

B. STATEMENT OF THE PROBLEM ……. 4

C. GOALS & AIMS……. 7

D. SIGNIFICANCE…….. 9

CHAPTER II REVIEW OF RELATED LITERATURE…….. 11

A. NEED AND BENEFIT OF CLINICAL DATAWAREHOUSE 11

B. SHORTCOMINGS OF RELATIONAL MODELS ……. 13

C. SEMANTIC WEB TECHNOLOGIES ……..… 13

D. PROPERTY GRAPH DATABASES…….…… 15

E. ACUTE KIDNEY INJURY……... 16

F. BAYESIAN INFERENCE…….…… 17

G. MARKOV CHAIN…… 18

CHAPTER III METHODS…….. 19

A. INTRODCUTION……..……. 19

B. RESEARCH DESIGN…….…….. 19

B.1. PROPERTY GRAPH MODEL…….. 19

B.2. PROBABILISTIC MODELS…….… 22

B.2.1. BAYES' THEOREM……. 23

B.2.2. OUTCOME RISK TOOL DEVELOPMENT…….. 25

B.2.3. MARKOV CHAIN……..…… 25

C. INSTITUTIONAL REVIEW BOARD (IRB) APPROVAL 27

D. DATA COLLECTION……. 27

CHAPTER IV IMPLEMENTATION…….…… 29

A. PROPERTY GRAPH MODEL IMPLEMENTATION AND DATA LOAD…….. 29

B. BAYES' THEOREM OUTCOME RISK USER INTERFACE 33

C. MARKOV CHAIN MODEL IMPLEMNTATION…… 34

CHAPTER V RESULTS……. 37

A. ACUTE KIDNEY INJURY GRAPH MODEL……..... 37

B. OUTCOME RISK ESTIMATION……. 40

C. NON-AKI TO AKI TRANSITION PROBABILITIES... 49

CHAPTER VI DISCUSSION…….……. 52

CHAPTER VII CONCLUSION…….……... 57

REFERENCES……..…… 59

APPENDICES……..……. 67

A. APPENDIX I – INSTITUTIONAL REVIEW BOARD (IRB) LETTER…….. 67

B. APPENDIX II – DEFINITIONS…….... 68

LIST OF TABLES

TABLE 1. STUDY COHORT BREAKDOWN……..…… 24

TABLE 2. AKI ICD-10 CODES…….…….. 28

TABLE 3. GRAPH DATABASE OBJECTS…….…….. 33

TABLE 4. SAMPLE GRAPH DATABASE QUERY…….. 39

TABLE 5. CALCULATED CLINICAL VARIABLES PROBABILITIES…… 43

TABLE 6. CLINICAL VARIABLES' STATISTICAL RELVANCE ……..… 56

LIST OF FIGURES

FIGURE 1. STAR SCHEMA STRUCTURE……..... 3

FIGURE 2. DATA LAKE……. 6

FIGURE 3. HYPOTHESIS-DRIVEN VERSUS DATA-DRIVEN MODEL……... 8

FIGURE 4. ANALYTIC COMPONENETS OF PROPOSED MODEL…….… 8

FIGURE 5. NOSQL DATABASES……..… 20

FIGURE 6. AKI WHITE BOARD DESIGN IN GRAPH DATABASES……..… 21

FIGURE 7. BAYES' THEOREM EQUATION…….…… 24

FIGURE 8. MARKOV CHAIN MATRIX……. 25

FIGURE 9. MARKOV CHAIN WITH INITIAL VECTOR PROBABILITIES… 26

FIGURE 10. AKI TRANISITION STATES WITH MARKOV CHAIN…… 26

FIGURE 11. AKI SCHEMA STRUCTURE……..….. 30

FIGURE 12. PROFILING OF AKI GRAPH DATABASE NODES…..…… 31

FIGURE 13. AKI RELATIONSHIPS AND COUNTS…… 32

FIGURE 14. AKI OUTCOME RISK USER INTERFACE……. 35

FIGURE 15. STATE TRNASITION PROBABILITIES…… 36

FIGURE 16. DISPLAY OF ALL CLINICAL EVENTS FOR A PATIENT…… 38

FIGURE 17. VISUALIZATION IN GRAPH DATABASES – SAMPLE QUERY 1 38

FIGURE 18. VISUALIZATION IN GRAPH DATABASES – SAMPLE QUERY 2 40

FIGURE 19. AKI PATIENTS' OUTCOMES IN CONTEXT OF CLINICAL VARIABLES……. 42

FIGURE 20. AKI PATIENTS' OUTCOMES RATES FOR EACH CLINICAL VARIABLE……. 42

FIGURE 21. SAMPLE OUTCOME PROBABILITIES USING BAYES' OUTCOME TOOL…….. 44-49

FIGURE 22. TRANSITION PROB. WITH INITIAL VECTOR FOR CLINICAL VARIABLES…….. 50

FIGURE 23. TRANSITION PROB. W/O INITIAL VECTOR FOR CLINICAL VARIABLES…… 50

FIGURE 24. TRANSITION PROBABILITIES FOR COMBINED CLINICAL VARIABLES…….. 51

FIGURE 25. SERUM CREATININE TRAJECTORY AS A CONTINEOUS VARIABLE…… 55

FIGURE 26. GRAPH SIMILARITY ALGORITHMS…….. 58

ABSTRACT

The increased adoption of electronic medical records (EMR) systems and emergence of clinical data warehouses to integrate data from diverse data sources energized clinical research and prompted the biomedical informatics community to envision and implement efficient and effective tools to facilitate conduct of research. Data warehousing, a valuable platform to provide clinical data for secondary use, is one tool, traditionally built using relational database models. Though relational models proved solid in data management applications across industries, the complexity and variety of clinical data require an agile technical environment that responds to evolving research data needs. A property graph model's data connectedness, data exploration, and visualization capabilities make it a solid candidate to represent and manage clinical knowledge. This study uses acute kidney injury (AKI) disease, an important and often overlooked disease process, to represent clinical data extracted from institutional data warehouse in a graph model. The resulting AKI graph model, which consists of entities (nodes) connected through meaningful relationships (edges), provides easy access to explore and view query results in either graphical or tabular format. The AKI model, conceptually a data lake, is horizontally scalable, which can integrate with other graph-based clinical domains of knowledge. Moreover, the AKI graph schema provides the right structure for a Bayesian network, which helps implement a Bayesian inference model to estimate AKI patients' outcomes probabilities, and also helps envision a Markov Chain transitions model to predict non-AKI patients' probabilities of requiring dialysis within a 48-hour.

Keywords: acute kidney injury, AKI, graph database, relational database, data warehouse, Bayes' theorem, Markov Chain

ACKNOWLEDGEMENT

First and foremost, it is God's willing helped me through completion of this dissertation. The continuing support and encouragement of my wife Mayada, and my children Amanda, Samantha, and Salma, who have repeatedly kept me on track to get where I am today. However, the continuing guidance and support of my advisor Dr. Shankar Srinivasan and the dissertation committee members, Dr. Suril Gohel's and Dr. Corey Hayes' valuable feedback was instrumental in helping me successfully achieve this milestone. My thanks to John Arthur, MD, PhD at the University of Arkansas for Medical Sciences (UAMS) for his clinical domain expertise in nephrology, especially in the management of acute kidney injury patients. Finally, my thanks to the University of Arkansas for Medical Sciences' Translational Research Institute (TRI) for the multi-centre collaboration opportunity that helped conceptualize my dissertation idea in the first place.

CHAPTER 1 INTRODUCTION

A. BACKGROUND OF THE PROBLEM

Adoption of clinical enterprise data warehouses (EDW) among academic medical centres has been on the rise with the aim to integrate data silos and improve data quality. The enormous amounts of data generated by electronic medical records (EMRs) and fragmented, heterogeneous patients' data incentivized the surge of data warehousing as the means to tap into the wealth of clinical data and aid both clinicians and management in driving evidence-based decision-making. Data warehouses are indispensable in providing valuable insights into ongoing clinical operations and monitoring of business trends in many healthcare organizations, and increasingly EDWs demonstrated their value in numerous clinical use cases, including monitoring antimicrobial resistance, measure antimicrobial use, and identify hospital-acquired infections [11]. They are also used to identify operational inefficiencies, including patient wait times and resource availability. While some organizations limit EDW use to fulfil reporting needs, others use as point of care decision support systems, help physicians consolidate and manage health information across the continuum of care, and effectively build and test disease and risk stratification predictive models.

The design and implementation of a clinical data warehouse is a major undertaking and requires organizational commitment and coordination that engages clinicians, clinical leaders, hospital service line staff, and information technology staff. The complexity of the United States' healthcare system is widely represented in institutional healthcare workflows, and is manifested in the adopted electronic medical record

(EMR) systems. The non-triviality of designing and implementing a clinical data warehouse that embodies the realities of a healthcare system led many healthcare organizations to opt-in to acquire vendor-based clinical data warehouses, while others were interested in implementing tailored EDWs that meet their data needs. In fact, vendor-based solutions are not without shortcomings. They provide a one-size fit all model that entails extensive retooling to match institutional expectations.

There are a number of design methodologies to implement a data warehouse, with the common goal of integrating disparate data sources. While early EDWs followed the relational entity-attribute-value model by normalizing data to the third normal form, known as corporate information factory, recent EDWs adopted a de-normalized design approach, aka dimensional or star model. Because queries running against a data warehouse can quickly grow in complexity, the star schema design provides faster query execution time and simplified query structure. Irrespective of the design methodologies of current day EDWs, the backbone of data structures and storage rely on relational database technology. Relational databases have been around since the 1970s and have proven to be solid in digitizing paper forms and automating well-structured business processes. The exponential growth of complex clinical data raised concerns about the efficiency and suitability of relational models in representing medical knowledge for the purpose of clinical research. While EDWs have been serving operational needs and patient quality projects well, there has been increased interest in repurposing clinical EDWs to meet organizational research data needs for the purpose of clinical and translational research.

The relational paradigm does not equate well with the expected, agile value promised by clinical enterprise data warehouses. The steps involved in designing a relational –based data warehouse include identifying, conceptualizing, and grouping related subject areas into relational database tables linked together by primary and foreign keys. Fig. 1 shows a dimensional star schema, where the observation, red coloured, table represents the facts and the green coloured tables represent dimensions. Fact and dimension tables are joined to produce query results and a join can get very lengthy and complex. The model fairs well in classifying different subject areas, patient, laboratory, medications, procedures, diagnoses, and providers, but falls short in representing definitions and relationships of data. The joins in a relational model are agnostic to the underlying semantic workflows that link each subject area with another.

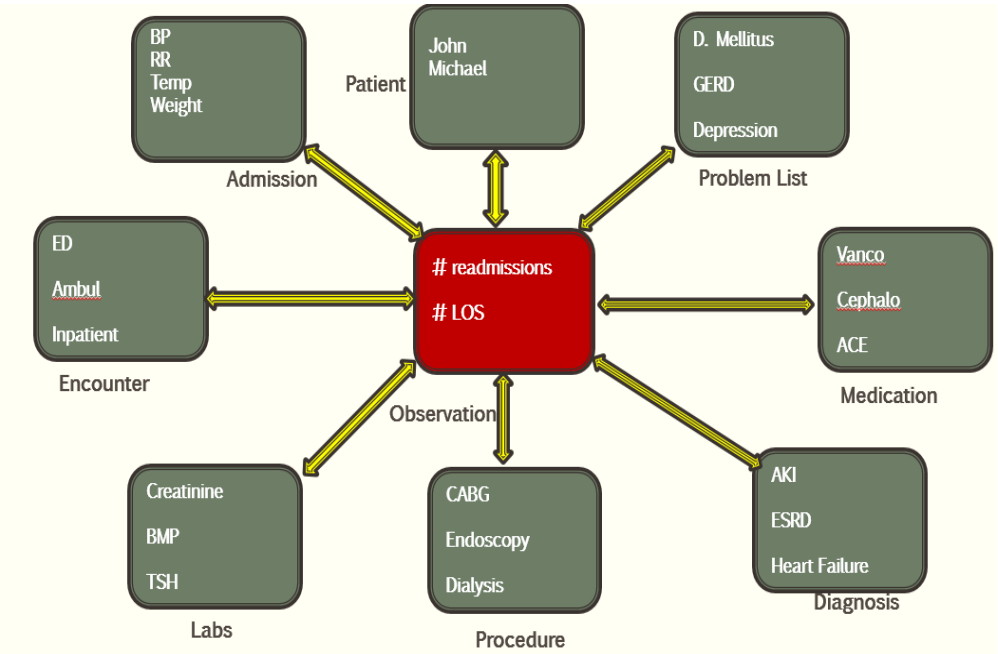

Fig. 1: Dimensional star clinical data warehouse. The observation, red colored, table is the fact table. The green colored tables are the dimensions that describe the facts.

B. STATEMENT OF THE PROBLEM

Electronic medical record (EMR) systems led to exponential, unprecedented growth of clinical data which motivated researchers to harvest for clinical and translational research. The design of EMRs facilitates charting individual patients' clinical information and is not well suited to link patients across disease groups, diagnoses, or demographics, and, therefore, is less effective in clinical research [27]. In addition, many healthcare organizations continue to maintain a large number of silo, ancillary systems that serve specific departmental needs. The introduction of data warehouses aimed to integrate disparate data sources into a single, coherent data in order to draw meaningful and improved insights. The integration of data sources helps improve data quality and quantity, and improve business processes [20].

While early clinical data warehouses were used mainly to fulfil operational and quality improvements initiatives, there has been increasing interest in secondary use of clinical data to answer clinical research questions and to discover new insights. The use of clinically derived data from electronic health records and other electronic clinical systems can greatly facilitate clinical research, and one approach for making these data available is to incorporate data from different sources into a joint data warehouse [3].

The relational model has been the de facto and backbone of a clinical data warehouse design and implementation, and building such a system is nontrivial. The process of deploying a functional clinical data warehouse that integrate data from one or more transactional systems requires, in addition to organizational commitment, identifying desired domain subject areas, designing a schema, and curating source data. The early-binding nature of the relational model makes it unsuited to accommodate the agility and scalability changing clinical workflows. Rapid advances in medicine and

evolving nature of healthcare systems require health information systems to be adaptable to new care standards and policies. The technical structure of a relational database model hinders data exploration [40], and lacks the necessary agility and scalability to adapt to changing data needs. Establishing strong, meaningful relationships among attributes in a relational model are not easy, resolving the many-to-many relationships are inefficient [44], and sorting through the variety of data types require reliance on highly skilled analysts to extract accurate and complete data.

The one-size fits-all relational, hypothesis-driven model is not capable of delivering the expected agility of a research data warehouse. The cycle of fulfilling research data requests is antiquated, time-consuming, and ineffective in competing for clinical research opportunities. Researchers are interested in having timely access to accurate data to answer research queries. Leveraging the right resources, including the right data model and self-service tools, ultimately results in large time and cost savings for the researcher as well as reproducible and accurate results [11].

I am proposing a flexible and horizontally scalable model that provides researchers the right platform to gain insight into acute kidney injury clinical domain by visually exploring patients' clinical facts, and querying and retrieving information ready, actionable results. The model is extendable and can grow to encompass new clinical domains without significant schema structure changes. This study hypothesizes that data connectedness in a graph model adds value through persistence of meaningful relationships between entities. An AKI connected graph model provides a complete network topology of AKI patients and their complete clinical profile, which allows clinicians and researchers have a 360-view of patients' recycling through healthcare system.

The proposed research data model is modular and can serve data needs of investigators interested in discovering knowledge in the area of acute kidney injury disease. The model is horizontally scalable and can easily scale up and integrate with other clinical domains with little or no modification to data structures. Conceptually, a graph model represents a data lake of interrelated entities, which can coalesce with other data lakes if there exists at least one relationship with another entity from adjacent data lakes, Fig. 2.

Another aim of the study is to shorten the research life cycle by moving compute time to the source through embedding advanced analytic methods within graph model to further gain insight.

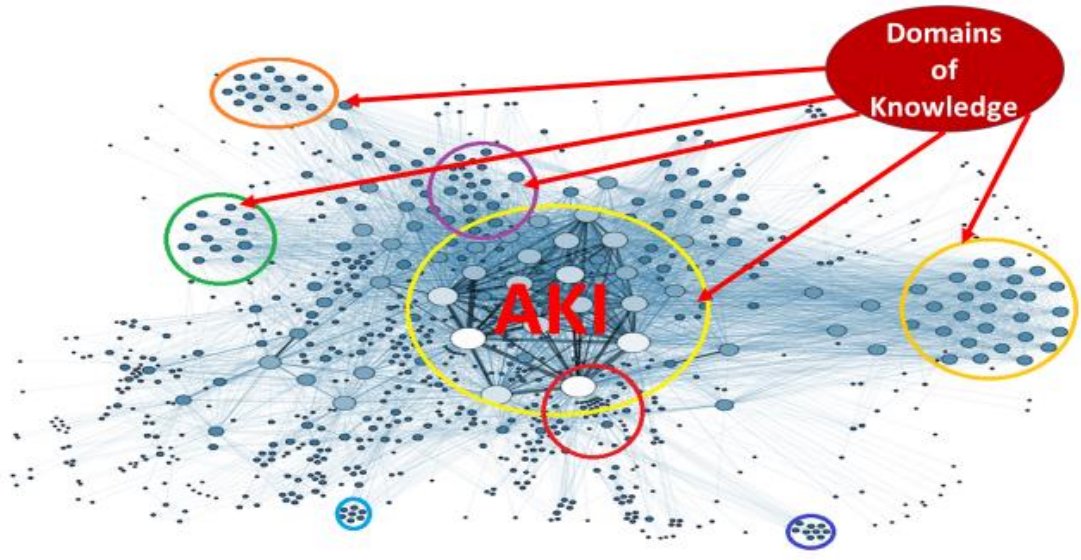

Fig. 2: Data lake – a set of semantically connected data elements make up the AKI data lake. The AKI data lake can intersect with another data lake when there exists a common semantic relation.

C. GOALS & AIMS

The long term goal of the research is to formalize and develop a semantically-infused, by establishing meaningful relationships, research data warehouse that inherently embodies relationships between various attributes to closely represent real world clinical workflows. The goal is to transform the relational model based UAMS clinical data warehouse from a hypothesis-driven model to a data-driven model.

The study aims to implement, using a property graph model, a horizontally scalable AKI data lake as a means to design future clinical research data warehouses. Specifically, the aims of the current work are three folds:

1. Design and populate a property graph model with acute kidney injury (AKI) clinical facts extracted from the University of Arkansas for Medical Sciences clinical data warehouse, Fig. 3.
2. Develop a predictive model to estimate AKI patients cohort probabilities of developing any of three outcomes of interest based on predetermined clinical variables, Fig 4. Design a user interface that allows the selection of clinical variables to estimate combined probabilities. Patients' outcomes of interest are:
   A) Alive and dialysis free: received dialysis, are alive, and no longer need dialysis.
   B) Alive and dialysis dependent: received dialysis, are alive, but still require dialysis.
   C) Dead: received dialysis, and died during or thereafter.
3. Conceptualize and construct a model to estimate likelihood of a hospitalized, non-AKI patient, that meets some or all predetermined clinical variables, requiring dialysis within 24 to 48 hours, Fig 4.

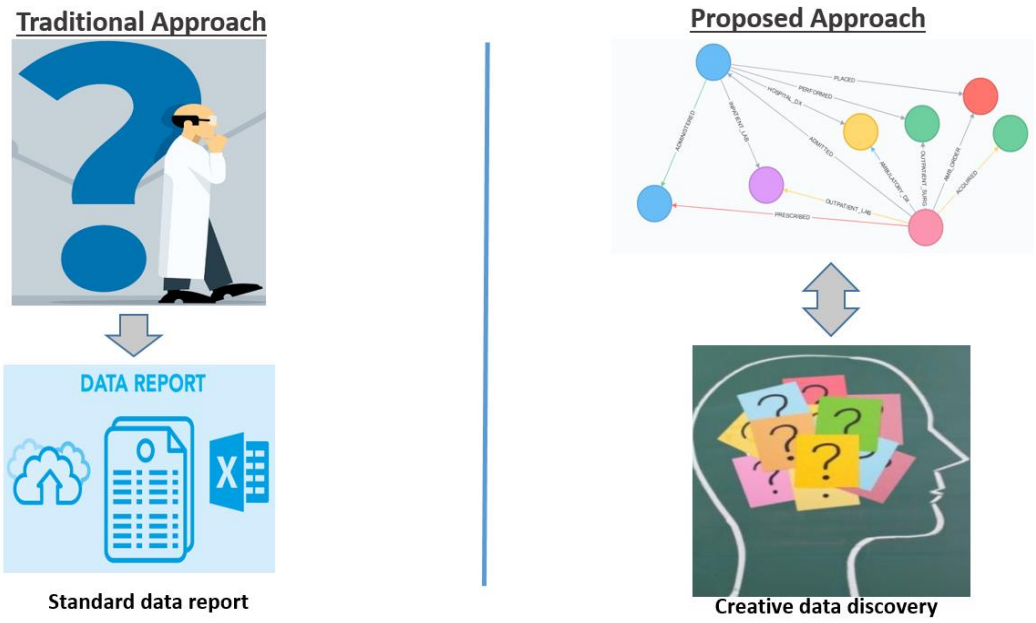

Fig. 3: Proposed model is to transition from hypothesis-driven model to data-driven, creative discovery model.

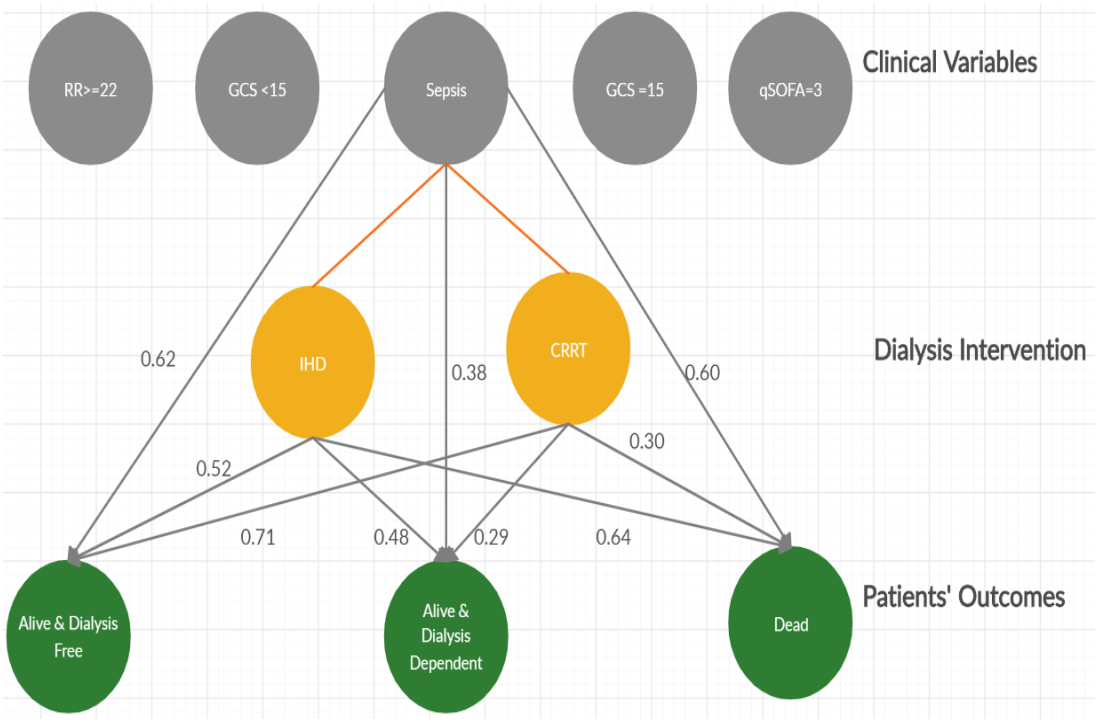

Fig. 4: Analytic components of the model. The Bayesian network/ inference model of patients' outcomes is indicated by the directed edges of the graph, along with their estimated probabilities. The Markov Chain state transition model is represented by the undirected edges to IHD (intermittent hemodialysis) and CRRT (continuous renal replacement therapy). For simplicity, the Bayesian and Markov model representations are shown for 'Sepsis' clinical variable.

D. SIGNIFCANCE

A. The Clinical and Translational Science Award (CTSA), through the National Institute of Health, motivated clinical research and, therefore, increased demand of clinical data by leveraging electronic medical record systems data and integrating silo data sources to fill emerging need. The current research lifecycle, from a research question conceptualization to data analysis, is labour intensive and lengthy. The process requires a researcher to work with an analyst to define study inclusions and exclusions, diagnoses, labs, and medications terminologies to identify a population cohort. The cycle may take weeks until an adequate data set becomes available, which is followed by a meticulous, time-consuming step of data preparation and harmonization. Simply, the current, underlying relational models that store clinical data do not meet the expected research demands of the CTSA.

B. The proposed model aims to shorten the cohort estimation and data extraction times of research projects lifecycle. The model accelerates clinical research by providing timely access to data that help researchers assess projects' feasibility quickly. Researchers will no longer have to wait weeks to determine whether to pursue or abandon a research idea. The proposed framework would provide instantaneous access to a study cohort of interest and can potentially allude knowledge that may not be easily discoverable otherwise.

C. The proposed model is proof of concept that provides a targeted platform for the conduct of research in specialized clinical domains. Researchers are not always

interested in having access to broad clinical data as is the case with present day clinical data warehouses; rather, they are more interested in having a research repository that provides access to actionable data. The proposed model is a standalone, a data lake, of AKI clinical domain, but not an isolated one. The AKI data lake would coalesce with another clinical domain, data lake, if there exists at least one shared relationship. The creation of a graph model allows researchers the opportunity to share their qualitative data in an innovative way that may provide new insight and enhance transparency of the analytical process by which they reach their conclusions [50].

D. Acute kidney injury (AKI) is a serious and highly heterogeneous disease, with variable links to poor outcomes and high mortality rate. AKI affects about 15% of hospitalized patients, but also has an incidence of 1% among the general population [79]. The etiology of AKI remains unknown but physicians rely on clinical phenotypes to diagnose. Identifying patients who are at higher risk of poor prognosis is still an enormous challenge in clinical practice [25]. Although current clinical classification of AKI severity depends on serum creatinine trajectory, there remains uncertainty in differentiating transient from persistent rises in serum creatinine. The high incidence and unknown etiology of the disease make AKI a very good candidate for proof of concept. The organization of AKI clinical facts in a property graph model provides investigators a new perspective on understanding the interplay of various clinical variables that influence AKI patients' outcomes.

CHAPTER II REVIEW OF RELATED LITERATURE

A. NEED AND BENEFIT OF CLINICAL DATA WAREHOUSE

A data warehouse in a healthcare organization has begun to show its capabilities not only in business operations, but also in clinical and translational research activities. Electronic health records are not very effective in clinical research and often not effective for translational research [27], because they are not designed nor optimized to link patients across a disease group, diagnosis, or patient demographic. The literature highlights the value of clinical data warehouses in integrating data from disparate data sources into a unified data repository, and the ability to disseminate data to researchers to conduct secondary use of clinical data as well as the ability to use for research studies and clinical trials cohort identification and study feasibility purposes. The use of clinically derived data from electronic health records (EHRs) and other electronic clinical systems can greatly facilitate clinical research as well as operational and quality initiatives, and one approach for making these data available is to incorporate data from different sources into a joint data warehouse [3].

The number of clinical use cases of a data warehouse is reported in literature including studies on recruitment for clinical trials, gene-disease association, and family health history data patterns, public health, trends in drug use, diabetes, epilepsy, infection surveillance, and medical errors [13]. Since data arrives from disparate sources, e.g., EMR systems, disease registries, a data warehouse is expected to integrate, and improve quality, of heterogeneous data. Integrating heterogeneous clinical data into a central data repository is considered a necessary step for clinical

research [53], and it is often necessary to accumulate patient data from several data sources in order to answer specific research questions.

Reference [20] highlighted the benefits of data warehousing on multiple levels, such as timesaving for users, improved quantity and quality of information, informed decision-making, and improved of business processes. Whereas [13] demonstrated the functionality and capability of a data warehouse in case studies, including multi-drug resistant pathogens and high risk of venous thromboembolism. In both use cases, the data warehouse served an important role in decision support and assisted clinicians in the early recognition of potential cases. In addition to disseminating rich data for the conduct of clinical research, data warehouses have emerged as a viable source for cohort identification to determine study feasibility.

Researchers are often unaware of the complexity in clinical data systems, how and when data are captured, and for what purpose. For instance, changes to diagnostic and billing codes, clinical unit names, laboratory test names, and other parameters occur over time; algorithms used to extract data on a cohort over time must therefore account for this complexity. Producing an optimal dataset often requires multiple iterations of cohort definition and algorithm refinement with clinical users, software development staff, and database administrators. Leveraging the right resources ultimately results in large time and cost savings for the researcher as well as reproducible and accurate results [11]. In fact, researchers are interested in having quick and accurate answers to their research questions through use of self-service tools.

B. SHORTCOMINGS OF RELATIONAL MODELS

In traditional relational database approaches, the technical structure often gets in the way of exploratory data analysis either by visualization or through data mining techniques [40]. To mitigate the limitation, linked data are explored by studies to integrate clinical and biomedical data using ontologies. Reference [48] demonstrated use of linked data to integrate Stanford's STRIDE clinical data warehouse into an integrated, semantic knowledge base that uses ontologies to bridge the gap between clinical and biomedical data, and the integration has the potential to accelerate translational research from the bedside to the bench via efficient approaches for knowledge discovery. The semantic clinical data warehouse, [48], that resulted from the integration represents a machine and human interpretable, formal knowledge representation that is much more expressive than a standard SQL-based clinical data warehouse. In spite of the successful integration into a linked data warehouse, scalability and query performance remain an issue, and there remains a need to develop more expressive and sophisticated queries that account for the semantics of hierarchies and terminologies. Reference [44] points out the limitations of relational database models in resolving the many-to-many relationships between genetic variants to individual relationships, where a query can become quickly inefficient as the number of relationships increase; moreover, relational database models lack of schema extensibility as a genetic variant is associated with increasing number of annotation sources.

C. SEMANTIC WEB TECHNOLOGIES

Linked data technology, also referred to NoSQL, emerged to represent complex medical knowledge and to allow flexible and scalable structure. Semantic web

technology, a NoSQL flavor and also known as graph database, relies on using Resource Descriptor Framework (RDF), or triple store, and ontology as the semantic layer. Resource Description Framework (RDF) and biomedical ontologies are having a strong impact on how knowledge is generated from biomedical data, by making data elements increasingly connected and by providing a better description of their semantics [5]. Semantic web technologies have been developed to overcome the limitations of the current web and conventional data integration solutions [43]. Reference [24] showed how the use of semantic relations instead of concept co-occurrences for literature-based discovery (LBD) are more natural and efficient. Reference [34] used the semantic web paradigm, instead of schema matching approaches (e.g. for relational databases), for the flexible representation of facts together with a semantic layer for describing corresponding types and relationships by ontologies (RDFS, OWL); and reference [47] presented a flexible, and scalable to a highly complex alternative to a relational model, using NoSQL, to store, retrieve and share ethnomedicinal plant data.

While linked data technology made headways, some studies tried to mix NoSQL with relational models. Reference [40] demonstrated that highly connected and sparse networks can be stored in a relational database, generally traversal-type queries, which connect data linked by different relationships, become too computationally expensive and cumbersome to design. Reference [39] developed OntoCRF, an ontology-based system using relational database to assist in speeding up clinical data repositories, shows that the approach is more flexible and efficient to deal with complexity and change than traditional systems; however, OntoCRF was not capable of managing explicit knowledge related to processes.

D. PROPERTY GRAPH DATABASES

A more recent newcomer to the NoSQL family and to the graph database is "property graph". In contrast to RDF, a property graph has nodes and edges, compact query syntax, and easy to read. A property graph database is inherently a "closed world" solution, it is possible to collect the entirety of meta-data about the types and number of entities and relationships between them [40]. Reference [9] highlighted that the purpose of data integration is to connect related data elements to enhance knowledge, and graphs represent the perfect cases for data integration purposes, where records represented as vertices are tightly connected together by the edges representing an equivalence relationship. Though RDF shows more flexibility in terms of API, building one for a graph is simple and intuitive. A great advantage is provided by the option of directly adding properties to nodes and edges. The main obstacle to design substructure search software with RDF has been the ontology definition [1]. Property Graphs were seen as ready-to-use solution for prototyping software; however, when performance and interoperability between different resources is considered, an RDF triple store appears as a more efficient technology [1]. Reference [4] used graph database to explore human metabolic data by envisioning particular metabolites together with their network neighborhood; Reference [32] described how graph theory helped ontology engineers understand ontology mappings such as how ontologies overlap and evolve, and to carry out tasks like finding new annotations, supporting other data integration methods, combining related ontologies, or ontology reuse. Reference [33] successfully used graph methodology to classify multiple sclerosis patients into different profiles using structural connectivity information, and offers new opportunities to identify potential biomarkers for the characterization of global as well as local effects of pathological mechanisms on brain networks. Reference [36] used network analysis, an application

area of graph theory, to visualize regionalized neonatal health care delivery systems and accurately depict known transport pattern. The approach supports the validity of network analysis as a tool to empirically quantify the degree of regionalization of neonatal care networks. Network analysis could thus be a powerful tool to define, analyze, and improve care at the network level in regions that lack a strong regulatory structure.

Property graph variant of a graph database has not been used to model clinical and patient specific data but was used to host large-scale biological interactions among genes, proteins, compounds and small RNAs. Reference [10] used property graphs to integrate a series of graph path search algorithms to discover novel relationships among genes, compounds, RNAs and even pathways from heterogeneous biological interaction data that could be missed by traditional SQL database search methods. However, reference [26] demonstrated that, though the semantic web technology has its place in importing and exporting data and metadata from a tumor model repository, graph models allow for storing metadata alongside model descriptions and ability to reason about relative scaling between parameters, removing much of the overhead that comes with systems built on RDF and OWL. In fact, reference [44] was able to store the Human Phenotype Ontology and the Gene Ontology in graph databases.

E. ACUTE KIDNEY INJURY

Acute kidney injury (AKI) is often an overlooked and unappreciated disease process that carries significant morbidity and mortality in up to half of critically ill patients [42]. AKI is a frequent complication of surgical procedures, and in Continuous-flow left ventricular assist devices implantation, AKI has incidence of 70% and is strongly

associated with death at 30 days and 1-year post-implantation [46]. AKI also has a 50% incidence rate in critically ill children; however, 98% of children with AKI survive their acute hospitalization and approximately 50% have longstanding subclinical effects on renal structure or function [52]. Hospitalized patients with severe AKI, requiring acute dialysis, have high rates of adverse outcomes during hospitalization and after discharge. The total number of death associated with dialysis-requiring AKI rose from 18,000 in 2000 to nearly 39,000 in 2009 [71]. In spite of the rapid rise in AKI incidence, research lags in understanding and discovering the reasons why some AKI patients recover, progress to chronic kidney disease, or die. Reference [71] created a logistic regression model using dialysis-requiring AKI as the outcome and a calendar year as the predictor. Reference [72] classified AKI patients into two sub-phenotypes, resolving and non-resolving, based on serum creatinine as a method to better define patients at risk for poor outcomes who might benefit from novel interventions.

F. BAYESIAN INFERENCE

Bayesian inference is application of Bayes' theorem to update a hypothesis as more evidence becomes available. It operates by combining prior belief or probability to predict posterior probability. Bayesian inference is used in wide range quantitative bioinformatics and clinical applications, including fMRI to detect magnitude change points and functional interaction patterns [75]. Reference [74] implemented a simple web interface that allow users to analyze a single dose-response data set, and also demonstrated Bayesian's approach outperformance over Marquardet-Levenberg algorithm. Key component of Bayesian inference is the prior and often times are difficult to quantify. Reference [77] highlights the importance of and the impact of priors, probabilities before observing new evidence, in the application of Bayesian

inference for latent biomarkers, and thus open an avenue for clinical implementation of new biomarkers. Reference [76] demonstrated improved accuracy of childhood diarrhea forecasts as generated by Bayesian inference in comparison to predictions made using only historical data trends.

G. MARKOV CHAIN

Markov models are useful when a decision problem involves risk that is continuous over time, when the timing of events is important, and when important events may happen more than once [70]. The ability of the Markov model to represent repetitive events and the time dependence of both probabilities and utilities allows for more accurate representation of clinical settings that involve these issues [70]. Reference [69] used precise Markov Chain to demonstrate the effect of air quality on patients' admissions, when air quality worsens transition probability from low-admission states to high-admission states increase dramatically. And, reference [78] used a decision tree model simulating hepatitis C virus screening and diagnosis was combined with Markov state transition model simulating treatment to evaluate cost effectiveness of a broad screening strategy for hepatitis C virus in the general population. The model calculated that more hepatitis C virus patients could be detected and treated with comprehensive screening compared to the current situation. Although Markov models represent one of the most common forms of decision-analytic models used in health care decision-making, correct implementation of such models requires reliable estimation of transition probabilities [68].

CHAPTER III METHODS

A. INTRODUCTION

This study aims to introduce a new semantic-infused research database model for acute kidney injury clinical area, modelled to represent clinical workflows of diagnosing, treating, and managing AKI patients in healthcare settings. Researchers can ask and receive answers to AKI related research questions through a user interface that empowers users by having quick access to data and ability to assess a research question feasibility. The inclusion of a semantic layer in the model can potentially provide inferences and insights about the data that may not be discernible otherwise.

B. RESEARCH DESIGN

Embedding semantics in a model requires careful design to represent a clinical domain such as acute kidney injury. Linked data offer effective solution to break down data silos; however, and creating such resources with sound characterizations of their meaning, using semantic annotations to common ontologies, is complex and requires human intervention [5]. Graph databases offer many features that make them an attractive option for a research-based setting where it might be necessary to dynamically develop and interactively mine heterogeneous data [40].

B.1. Property graph model

A NoSQL data model can be any of four types: key-value, column-family, document, or graph, Fig. 5. Graph models, property graphs, and resource description framework (RDF) are characterized by embedding semantics within their structures. RDF uses a triple store, in the format of subject -> predicate-> object, and an ontology

to add context and meaning. The RDF model requires clinical domain expert and ontologist to model underlying semantics. A property graph model, however, is flexible and representative of actual clinical workflows. It still requires understanding of a subject area and input of a domain expert to add meaning and context. Property graph consists of nodes (entities) linked by edges (relationships). Both nodes and edges can have properties that describe the nodes and edges. Model design in a property graph model follows whiteboard design; that is, entities and relationships generated during brainstorming sessions represent actual physical structures when database is created. Modelling of data in a graph database ideally are guided by the same principles as those used for ontology design, [40], and a Neo4j graph can be thought of as a collection of instances of data, where node "labels" are equivalent to classes, and types of edges- to relationship types.

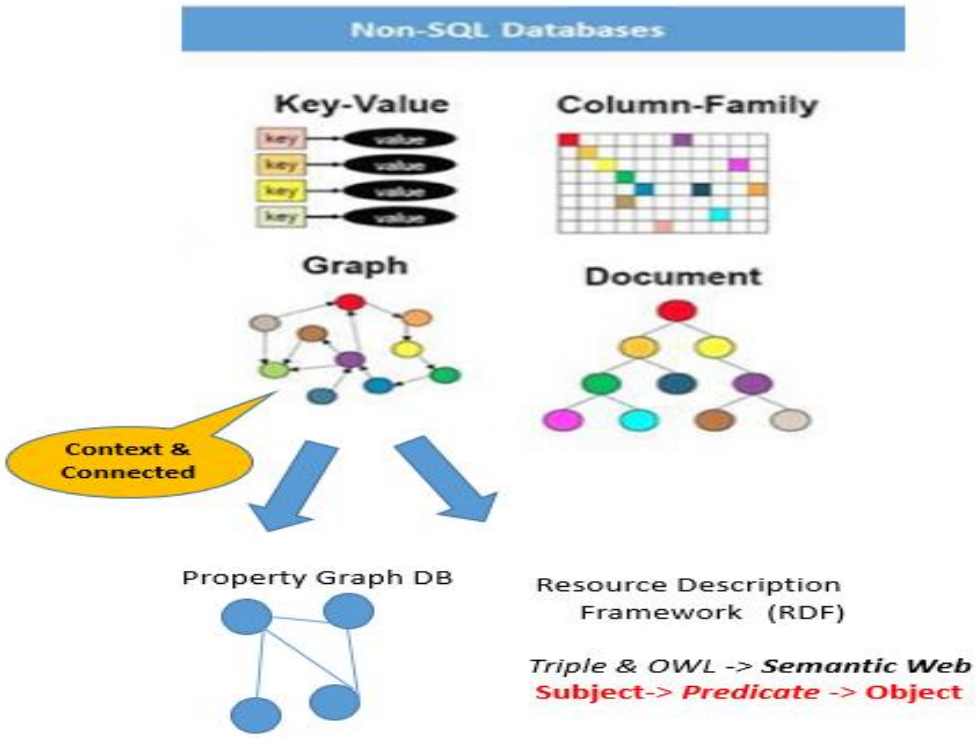

Fig. 5. Flavours of NoSQL data models. Graph models can be either property graph or resource description framework (RDF). RDF normally requires ontologies to add semantics. Property graph semantics are represented by adding meaningful relationships.

The whiteboard design of AKI property graph model, Fig. 6, describes patients' recycling within a healthcare system, where entities (nodes) can be person, disease, or location, and relationships (edges) represent relationships between entities. It is a property graph because properties because nodes and edges may contain attributes that further describe them. The model, Fig. 5, represents UAMS AKI patients' navigation of the healthcare system while being managed for acute kidney injury. It is worth noting that the proposed property graph model is not a replacement of a data warehouse; rather, it complements it [40]. The source of data for this study is the institutional clinical data warehouse; complete acute kidney injury patients' clinical facts are extracted and loaded into a Neo4j graph database.

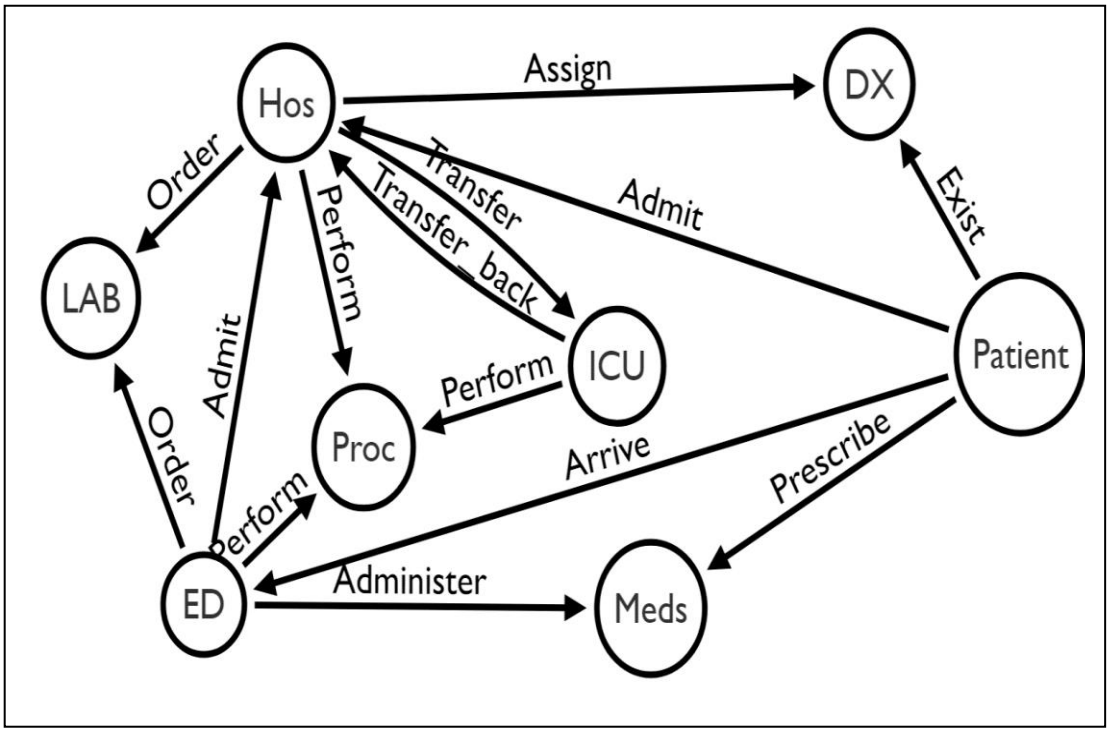

Fig. 6. Whiteboard design of Acute Kidney Injury patients' navigation of a healthcare system. A patient can be admitted to hospital through a scheduled appointment for a procedure or the emergency department. A patient may have a set of prescribed medications or existing clinical problems. Once in the healthcare system, a set of diagnostic workups are conducted, and depending on clinical presentation and laboratory results, a patient suspected of having AKI is started on dialysis, and also may be administered medications and fluids, and severely ill patients may be transferred to the intensive care unit (ICU).

B.2. Probabilistic Models

Possible patients' outcomes of acute kidney injury are dialysis free, dialysis dependant, or dead. The clinical variables that potentially impact any of the outcomes are identified from both acute kidney injury domain expert and validated from analysing the AKI cohort data set.

A) Sepsis: presence of infection diagnosis, hypotension, and fever. There are several diagnostic criteria such as lactate and increased heart rate to list a few. Sepsis associated AKI remains an important concern and a clinical burden in development of acute kidney injury [73].

B) Respiratory Rate (RR): normal RR is between 12 -22. Less than 12 or higher than 25 is considered abnormal. Respiratory rate greater than or equal to 22 breaths per minute impacts qSOFA (Quick Sequential Organ Failure Assessment) score and therefore AKI patients' outcomes.

C) Systolic Blood Pressure (SBP): pressure exerted when blood is ejected into arteries – normal SBP is 120 or less. Systolic blood pressure less than or equal to 100 mg/Hg impacts qSOFA and therefore AKI patients' outcomes.

D) Glasgow Coma Score (GCS): most common scorning system to measure altered mentation after traumatic brain injury, and is measured using three functions: eye opening, verbal response, and motor response. Scores range from severe (8 points or less), moderate (9-12 points), and mild (13- 15 points). AKI patients with GCS < 15 tend to have unfavourable outcomes.

E) qSOFA: a bedside assessment to identify patients at greater risk for poor outcomes (mortality) outside the ICU. It uses three criteria: one point for SBP < 100mmHg, one point for RR >=22 Breaths per minute, and one point for GCS < 15. Scores range from low risk of mortality (score = 0) to high risk of in-hospital

mortality (score= 2-3). A score of 2 or 3 negatively impacts AKI patients' outcomes.

F) FR (Fluid Removal): the volume of fluid removed is clinically believed to impact AKI patients' outcomes.

The following two dialysis modalities are not measurements of clinical variables; rather, a patient being on one or the other signifies that the level of illness severity versus the other, and therefore could impact a patient outcome.

G) Intermittent Hemodialysis (IHD): dialysis modality AKI patients receive with the goal to normalize kidney function.

H) Continuous Renal Replacement Therapy (CRRT): dialysis modality provided to severely ill AKI patients in intensive care units (ICU).

To develop a model that calculates patients' outcome probabilities – alive and dialysis free, alive and dialysis dependent, and dead, Bayes' theorem is used to calculate conditional of each and combined probabilities of patients characteristics.

B.2.1. Bayes' theorem: Fig. 7, consists of *prior probability*, which is an initial probability value obtained before any additional information is obtained, and a *posterior probability*, which is a probability value that has been revised by using additional information that is later obtained. In preparation to estimate conditional probabilities, Table 1 is produced from querying the property graph database. The table lists 13 clinical variables and their corresponding breakdown counts of patients for each of the outcome variables of interest from the set of the 798 AKI patients.

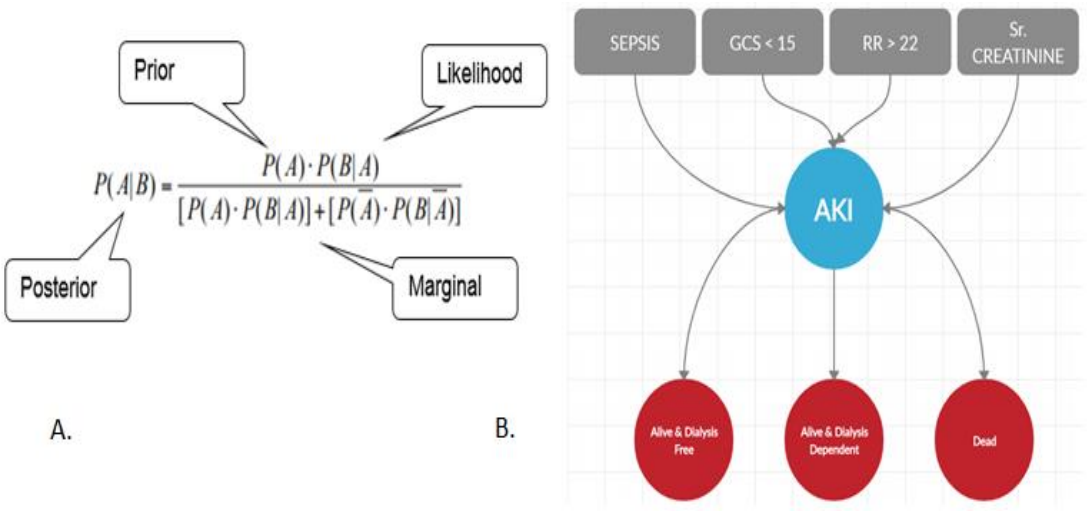

Fig. 7: A. Bayes' theorem: Posterior is probability given all observed evidence; likelihood is how probable the evidence given hypothesis is true; Prior is probability before seeing the new evidence; Marginal is probability of evidence under all possible hypotheses. B. Some clinical variables (Sepsis, GCS, RR, Sr. Creatinine) that play part in developing acute kidney injury (AKI). Patients with these clinical variables play role in impacting patients' AKI outcomes: Alive & Dialysis Free, Alive & Dialysis Dependent, and Dead.

Table 1
Study Cohort Breakdown

Clinical Variable	Alive & Dialysis Free	Alive & Dialysis Dependent	Dead	Totals
Total	243	159	396	798
IHD	119	109	100	328
CRRT	124	50	296	470
SEPSIS	73	44	169	286
GCS=15	110	77	82	269
GCS<15	133	82	314	529
RR < 22	168	116	202	486
RR >=22	75	43	187	305
SBP >100	195	130	272	597
SBP <=100	48	29	124	201
Qsofa=0	80	59	32	171
Qsofa=1	85	51	155	291
Qsofa=2	63	44	154	261
Qsofa=3	15	5	54	74
FR<1L	48	25	52	125
FR>1L	44	37	46	127

Note. Provides counts of various clinical variables in the context of outcomes of interest for the study of acute kidney injury patients. It is already obvious that patients with qSOFA=3, for example, have high risk of mortality in respect to the total of patients with the same score.

B.2.2. Outcome Risk Tool Development

In order to provide a quick and easy method of estimating probability of each outcome, a web-based tool will be developed to show case its utility. The input file, Table 2, would be refreshed weekly to update evidence - prior probability. A user can select one or more clinical variables and probability risk for each of the three outcomes is produced. Outcome probabilities are dynamically updated as clinical variables are selected or deselected.

B.2.3. Markov Chain

Health technology assessment (HTA) and medical decision-making, more generally, rely on the use of decision-analytic models, [68], and Markov models are a popular form of decision-analytic models which characterize patient cohorts based on a finite number of mutually exclusive and exhaustive "health states". Markov chains provide intuitive method to statistically model random processes, and consist of a set of memoryless states, and state transitions are determined by probability distribution. Fig. 10 has two states: HEALTHY (non- AKI) and AKI, and transitioning from one state to another is defined by a transition matrix, Fig. 8, where the sum of each row adds to 1.

$$P = \begin{pmatrix} 0.998 & 0.002 \\ 0.480 & 0.520 \end{pmatrix}$$

Fig. 8. Markov Chain transition graph is mathematically represented by matrix. Based on the AKI cohort (n=798) definition for this study, the probability of developing AKI for all patients (n=430k) in the institutional data warehouse is 0.002.

A Markov Chain fact states that the power k of the matrix P represents the (i, j) probability to arrive from state i to state j at k steps. The fact is used to calculate probability of a patient transitioning from HEALTHY state to AKI state (needing

dialysis) at some time *k* in the future, Fig. 9. However, a future state may be impacted by the clinical variables, e.g, sepsis, GCS, qSOFA, or FR, initial probabilities, Fig. 10.

$$P = \begin{pmatrix} 0.998 & 0.002 \\ 0.480 & 0.520 \end{pmatrix}^{k} * (q1, q2, q3)$$

Fig 9. Markov Chain probability matrix of transitioning from either HEALTHY state to AKI state or vice versa after *k* steps. A patient has some initial probabilities (q1, q2, q3) for (SEPSIS, GCS, qSOFA), respectively. *k* steps may represent *k* lab tests or time intervals.

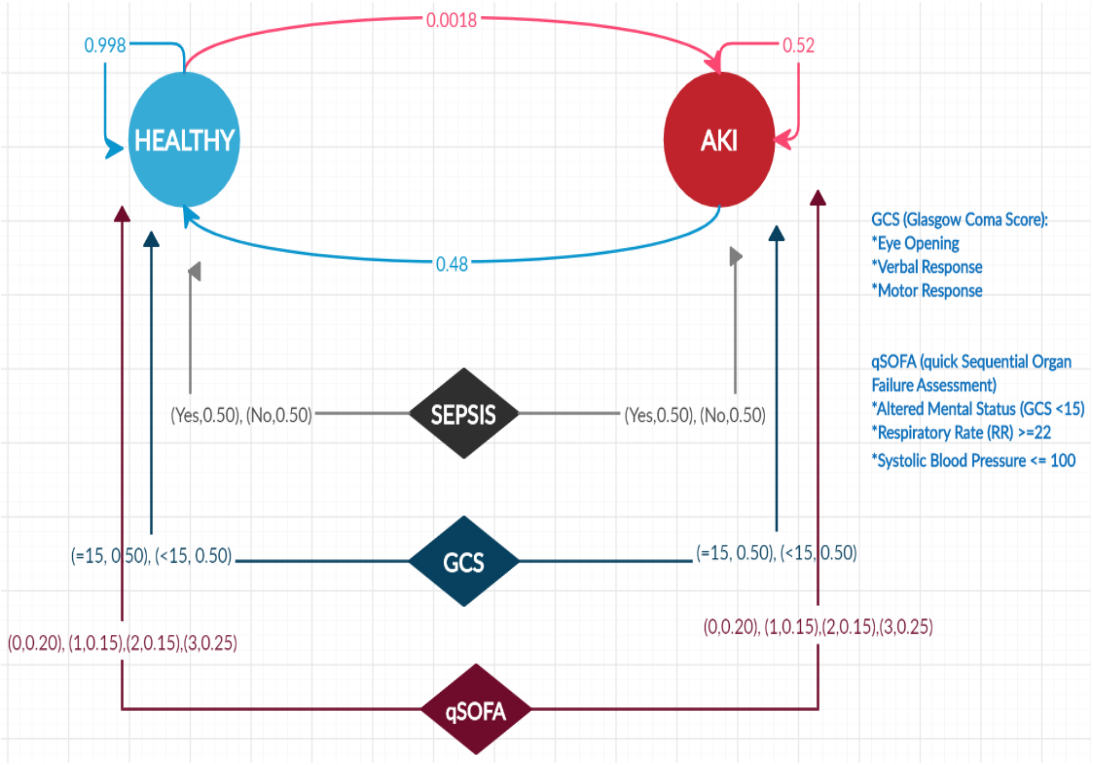

Fig. 10. Describes the two states Markov Chain for acute kidney injury. A patient can be in either HEALTHY or AKI states and transition back and forth in what is represented in a transition matrix, Fig 4. A set of scalar probabilities, SEPSIS, CGS, or qSOFA, impact whether a patient transition from one state to another. *Note*: The probabilities shown in the diagram are for demonstration purposes.

C. INSTITUTIONAL REVIEW BOARD (IRB) APPROVAL

In anticipation of requesting historical clinical data for secondary use from the institutional clinical data warehouse, a study protocol submitted to the UAMS' IRB (# 228146) and determined a non-human subject research, and approved on 05/24/2018. Copy of the IRB is in Appendix 1.

D. DATA COLLECTION

The property graph database is created from data extracted from the EDW at the University of Arkansas for Medical Sciences. The AKI patient cohort is determined as follows:

1. Patients meeting inclusion AKI diagnosis criteria, Table 2, as specified by ICD-10 (International Classification of Diagnoses-10)
2. Exclusion criteria include those with chronic kidney disease stage 5 (code N18.5), end stage renal disease (code N18.6), and kidney transplant patients (codes: T86.xx). The reason for excluding these diagnoses is to avoid bias in performing AKI outcomes analyses. Patients having these diagnoses do not recover and their conditions are terminal.

The types of data extracted from the institutional clinical EDW include: patient demographics, diagnoses, hospital admissions, vital signs at admission and throughout hospital stay, emergency department visits, comorbidities, laboratory results, prescribed and administered medications, medical and surgical procedures, problem lists, dialysis procedures detail, ventilation data, and Glasgow coma scale score.

Table 2
AKI ICD-10 Codes

Code	Coding system	Description	Entity type	List name
N14	ICD-10	Drug- and heavy-metal-induced tubulo-interstitial and tubular conditions	diagnostic	res30: Acute kidney injury
N14.1	ICD-10	Nephropathy induced by other drugs, medicaments and biological substances	diagnostic	res30: Acute kidney injury
N14.2	ICD-10	Nephropathy induced by unspecified drug, medicament or biological substance	diagnostic	res30: Acute kidney injury
N17	ICD-10	Acute renal failure	diagnostic	res30: Acute kidney injury
N17.0	ICD-10	Acute renal failure with tubular necrosis	diagnostic	res30: Acute kidney injury
N17.1	ICD-10	Acute renal failure with acute cortical necrosis	diagnostic	res30: Acute kidney injury
N17.2	ICD-10	Acute renal failure with medullary necrosis	diagnostic	res30: Acute kidney injury
N17.8	ICD-10	Other acute renal failure	diagnostic	res30: Acute kidney injury
N17.9	ICD-10	Acute renal failure, unspecified	diagnostic	res30: Acute kidney injury
N19	ICD-10	Unspecified kidney failure	diagnostic	res30: Acute kidney injury
N99.0	ICD-10	Postprocedural renal failure	diagnostic	res30: Acute kidney injury
R34	ICD-10	Anuria and oliguria	diagnostic	res30: Acute kidney injury
R94.4	ICD-10	Abnormal results of kidney function studies	diagnostic	res30: Acute kidney injury

Note. ICD-10 codes used to determine AKI cohort.
https://clinicalcodes.rss.mhs.man.ac.uk/medcodes/article/30/codelist/res30-acute-kidney-injury/

CHAPTER IV IMPLEMENTATION

A. PROPERTY GRAPH MODEL IMPLEMENTATION AND DATA LOAD

The entities (nodes) and edges (relationships) in a property graph follow a whiteboard design. The implementation and the physical representation of the objects follow the whiteboard design, Fig. 6, which represents actual entities (nodes) and the associated relationships (edges) for acute kidney injury patients, Fig. 11. A node label, e.g., Patient, refers to any of 798 unique AKI patients; a node may have a set of properties that describe it, e.g. Patient node has unique medical record number (MRN), demographic information, and patient disposition, e.g. dead. A relationship (edge) also may have a set of properties that further describe the relationship, e.g. Admitted relationship contains date of hospital admission for a patient encounter, since a patient may have one or more hospital admission. A relationship may be self-referencing, e.g., Temperature or Pulse in Inpatient node; meaning, the relationship (Pulse) is generated and measured during same hospital encounter. A relationship may also be bi-directional that indicates a patient transfer from an inpatient unit to the intensive care unit (ICU), and moves back to the inpatient unit, e.g. Transferred and Transferred_back relationships. Uniqueness of both entities (nodes) and relationships (edges) is enforced by creating constraints, e.g. MRN property in Patient node. Table 3 highlights counts of sample graph database objects, nodes and edges, created to represent acute kidney injury.

Data profiling is an important component of ensuring data quality of a database. Since a graph database overall well-being depends on establishing meaningful

relationships (edges) between entities (nodes), data profiling is critical to identify hidden patterns and improve understanding of metadata, Fig. 12, and Fig.13.

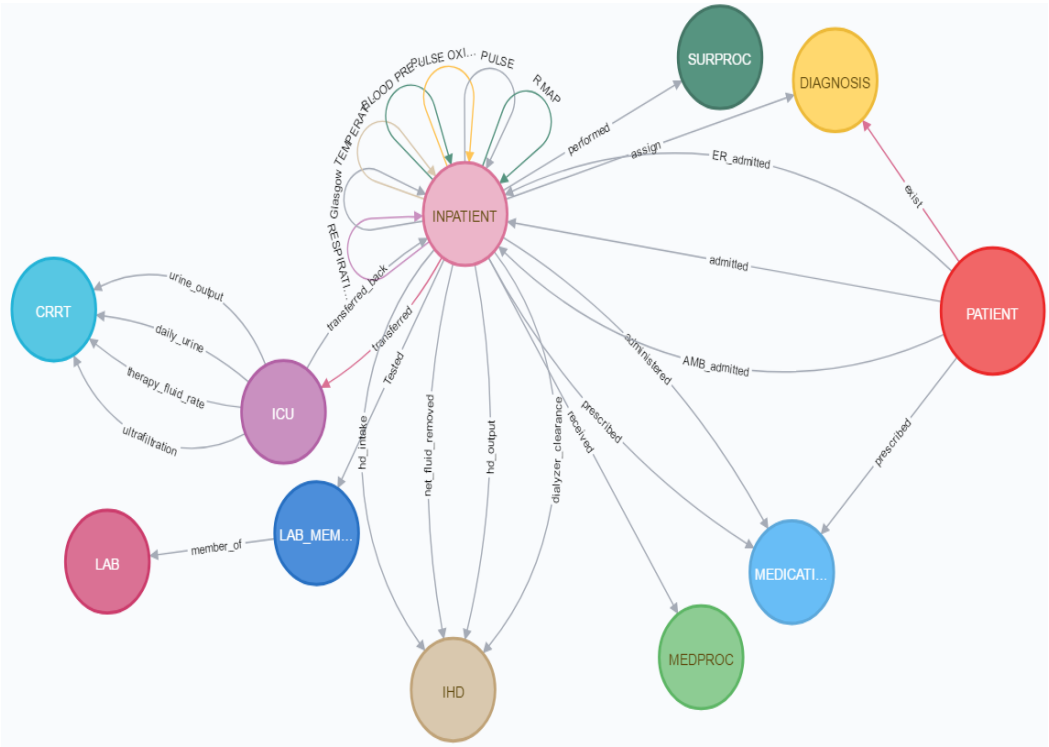

Fig. 11. The schema is patient-centred and all entities (nodes) and relationships (edges) reflect AKI patients' interactions within a healthcare system. The figure represents the actual layout of database objects and potential relationships between entities.

```
MATCH (n) WHERE rand() <= 0.1
RETURN DISTINCT labels(n) AS Nodes, count(*) AS Sample, avg(size(keys(n))) as Avg_Property,
min(size(keys(n))) as Min_Prop_CNT, max(size(keys(n))) as Max_Prop_CNT,
avg(size( (n)-[]-() ) ) as Avg_Rel_CNT, min(size( (n)-[]-() ) ) as Min_Rel_CNT,
max(size( (n)-[]-() ) ) as Max_Rel_CNT
```

Nodes	Sample	Avg_Property	Min_Prop_CNT	Max_Prop_CNT	Avg_Rel_CNT	Min_Rel_CNT	Max_Rel_CNT
["PATIENT"]	91	13.560439560439557	13	14	25.89010989010989	3	136
["DIAGNOSIS"]	1135	2.0	2	2	40.49779735682821	0	7746
["MEDICATION"]	403	2.0	2	2	185.10173697270463	0	21462
["INPATIENT"]	281	7.0	7	7	1054.1459074733096	9	22378
[]	37390	0.0	0	0	0.0	0	0
["ICU"]	162	3.2530864197530875	1	7	96.91975308641979	0	1865
["CRRT"]	97	0.845360824742268	0	2	116.16494845360826	0	1111
["IHD"]	61	2.0	2	2	19.836065573770487	4	85
["MEDPROC"]	136	2.0	2	2	114.90441176470588	0	10274
["SURPROC"]	31	2.0	2	2	1.5806451612903225	0	11
["LAB"]	88	1.0	1	1	5.784090909090912	1	97
["LAB_MEMBER"]	214	1.0	1	1	555.6869158878502	1	15445

Fig. 12. A profiling query that samples nodes and their statistics.

MATCH ()-[r]->() RETURN type(r) as RealtionType, count(*) as Total order by Total desc;

Relationship Name	Count
"Tested"	1658972
"administered"	1656676
"assign"	786606
"RESPIRATIONS"	215820
"PULSE"	196985
"PULSE OXIMETRY"	184911
"BLOOD PRESSURE"	171184
"received"	166396
"TEMPERATURE"	144500
"R MAP"	135175
"urine_output"	111912
"Glasgow"	72269
"therapy_fluid_rate"	71350
"ultrafiltration"	66658
"exist"	43038
"prescribed"	35154
"daily_urine"	32366
"hd_intake"	10712
"member_of"	7322
"hd_output"	4932
"net_fluid_removed"	4592
"dialyzer_clearance"	4458
"admitted"	4398
"transferred"	3170
"transferred_back"	3170
"performed"	2814
"ER_admitted"	1242
"AMB_admitted"	302

Fig. 13. Query list relationships and counts. The ‘test’ relationship is largest in reference to lab tests performed, followed by relationship ‘administered’ in reference to medications.

B. BAYES' THEOREM OUTCOME RISK USER INTERFACE

Bayesian probability represents a level of certainty relating to a potential outcome, and tries to quantify the trade-off between various decisions. A user interface was developed for this study to calculate the Bayesian probabilities of AKI patients' outcomes. Fig. 14 shows a screen shot of the actual user interface that produces outcomes' conditional probabilities estimates within the space of selected clinical variables. Related variables are grouped together and only one selection can be checked from a group, except for 'Dialysis Type', which either or both selections can be checked – a patient may receive both dialysis modalities during a hospital admission.

In preparation for designing and implementing AKI's outcome risk estimation tool, AKI patients' data are analysed and categorized, Fig. 17 and Table 5, in the context of the three outcomes of interest. The prior probabilities would be updated, by refreshing data from source, as new evidence becomes available.

Table 3
Graph Database Objects

Graph Database Object	Nodes	Relationships
Graph Data Model Schema	10	28
Number of Patients	798	27815
Total Inpatient admissions	2973	2891429
Total Medications	3845	845915
Total (IHD)	632	12347
Total CRRT	1117	141143

Note. Describes counts of acute kidney injury database data structures created to represent the 798 AKI patients. The 10 nodes and 28 relationships represent a high-level schema structure; there were ~3K patient admissions and about ~3million relationships; ~4K medication nodes and ~900K relationships; 632 IHD (intermittent dialysis) nodes and ~12K relationships; and ~1K CCRT (continuous renal replacement therapy) nodes and ~140K relationships.

C. MARKOV CHAIN MODEL IMPLEMENTATION

Markov models are useful when a decision problem involves risk that is continuous over time, when the timing of events is important, and when important events may happen more than once [70]. The Markov model is particularly useful in analyzing risk factors in cohort studies and has been applied successfully to the study of lung cancer and HIV infection [69]. The correct implementation of Markov models requires reliable and robust estimation of transition probability matrices (TPMs), and in the simplest case, transition probabilities can be estimated in a straightforward manner using nonparametric methods based on observed counts for movements between health states in a data source [68]. The study aims to study effects of clinical variables such as GCS, qSOFA, respiratory rate, and systolic blood pressure on development of acute kidney injury by using Markov Chain to predict probability of transitioning from a no-AKI (HEALTHY) state to AKI state. A Markov model assumes that a patient is always in one of a finite number of discrete health states, called Markov states. In this study, we are interested in the probability of a non-AKI (HEALTHY) state patient developing, by transitioning to, AKI state over a 24-48 hour period.

In retrospective, case-control study, the process of determining the Markov Chain transition probabilities for each clinical variable is determined by calculating the odds ratio and converting it to a probability. Again, we are mostly interested in knowing the probability of transitioning from a Non-AKI state to AKI state in 24-48 hours, but not vice versa. However, the Bayesian outcome risk tool is used to complete the state transition probabilities in the direction from AKI to non-AKI.

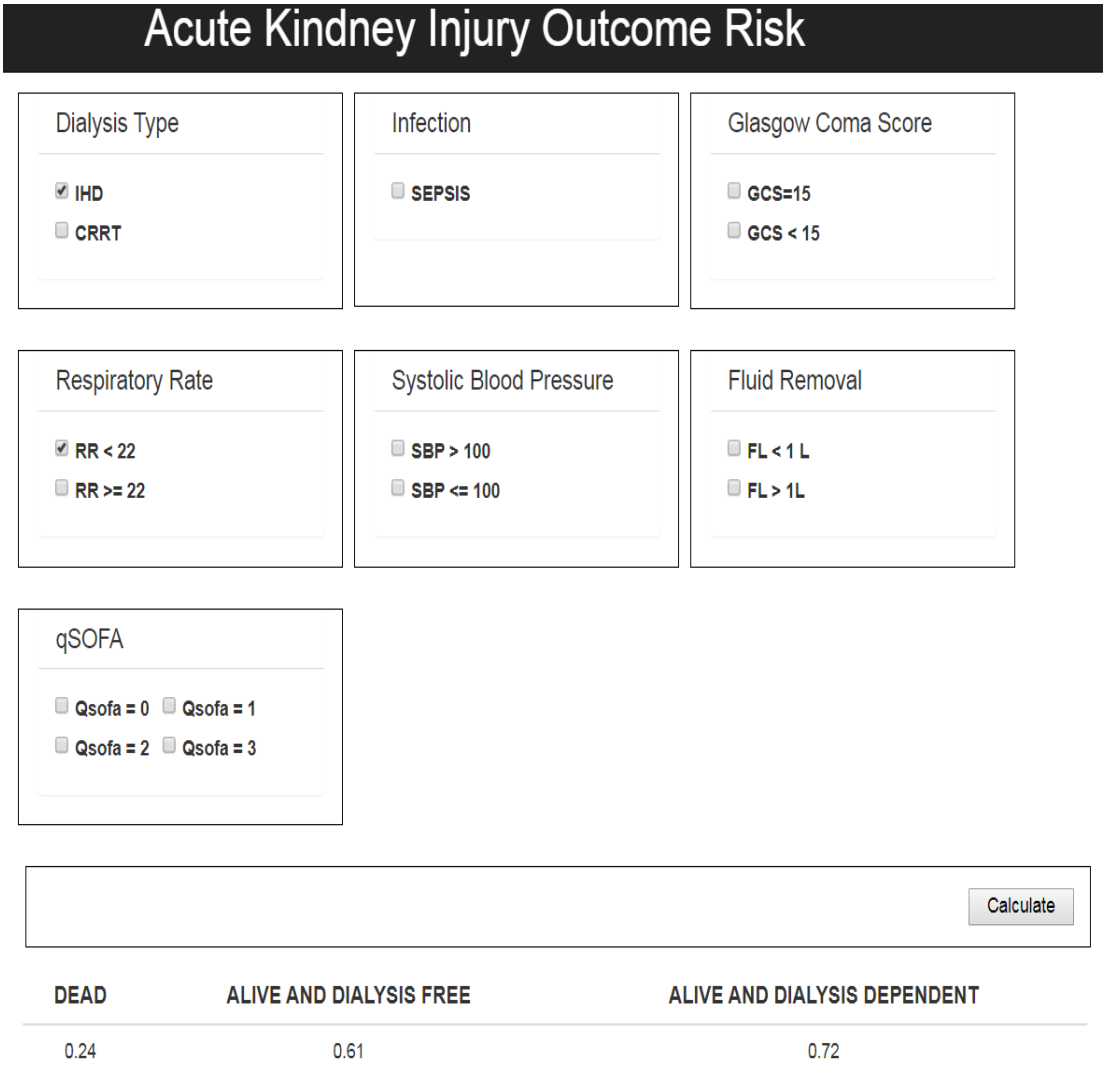

Fig. 14. User interface to estimate AKI outcomes risk. The tool operates by dynamically calculating Bayesian conditional probabilities based on clinical variables selections. Each box constitutes possible selections for each clinical variable of interest. A user may select either IHD, CRRT or both from the 'Dialysis Type' as a patient may receive both during a clinical encounter, depending on severity of illness. The web-based interface was developed using Microsoft ASP.NET framework and codes was written using Visual Basic.

The process of determining initial transition probabilities is described in Fig. 15. For each clinical variable, a 2x2 contingency table is generated, and each cell includes patient counts as determined from both the extracted AKI data set and data obtained from UAMS' clinical enterprise data warehouse.

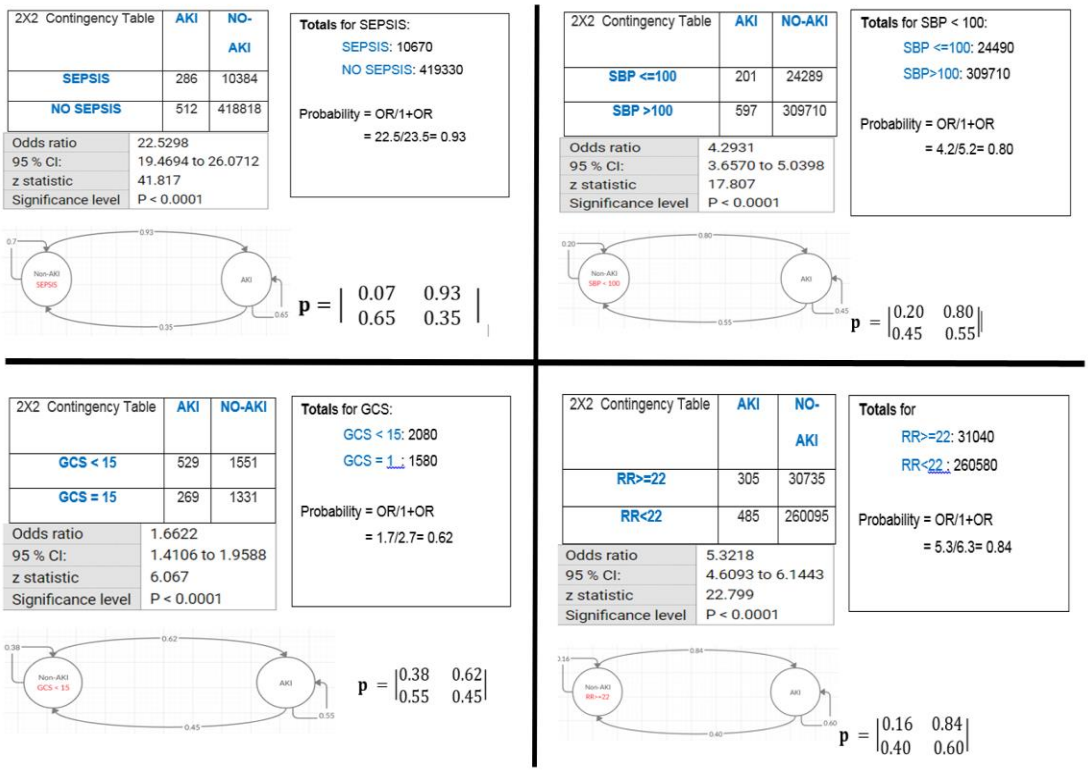

Fig. 15 State transition probabilities for four clinical variables of interest: sepsis, GCS <15, SBP <=100, and RR >=22. For each clinical variable, a 2x2 contingency table is shown, and displaying occurrence of AKI or non-AKI patients in respect to the total number of patients that meet, and not meet, the criteria from the institutional clinical data warehouse. The 2x2 contingency tables' values are from retrospective data, and are therefore case-control cases. The odds ratio for each clinical variable is calculated in order to derive the transition probability for each.

CHAPTER V RESULTS

A. ACUTE KIDNEY INJURY PROPERTY GRAPH MODEL

The designed and implemented graph model of AKI, Fig. 11, represents the schema, encompassing all related clinical facts within the AKI clinical domain. It is possible to display and view AKI's graph database objects at the same time; however, it is more feasible to work with subsets of nodes and relationships that share some characteristics to allow for data exploration and appreciate depth of interaction between entities and corresponding relationships. Fig. 16 displays a patient's clinical events during several hospital admissions, and Fig. 17 demonstrates an AKI patient recycling through UAMS' healthcare system. A clinician, by expanding nodes or relationships of interest, can gain valuable information about specific clinical events. A user can also quickly view a patient's assigned diagnoses, prescribed medications, hospital admissions, and medical and surgical procedures.

Property graph models are well suited to display query results graphically, excellent medium for data exploration. There are use cases that are beneficial to display query results in tabular format for either reporting or data analysis purposes. Table 4 displays an example output result of a query that lists in descending order length of hospital stay, dialysis free days, and name of clinical procedure.

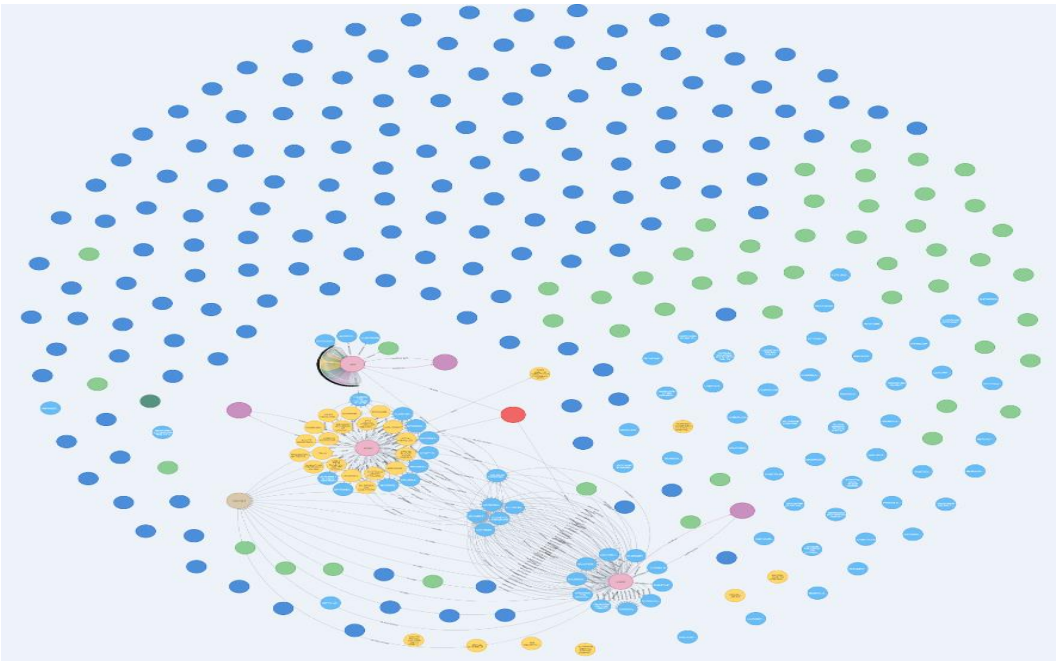

Fig. 16. A query that returns a patient's clinical events during hospital admissions. The patient (red node) has 3 hospital admissions (pink nodes), had 1 IHD (brown node) during an admission, was moved to ICU (purple node) 3 times. Medical and surgical procedures (green nodes), medications (light blue node), lab results (dark blue node), and diagnoses (yellow node)

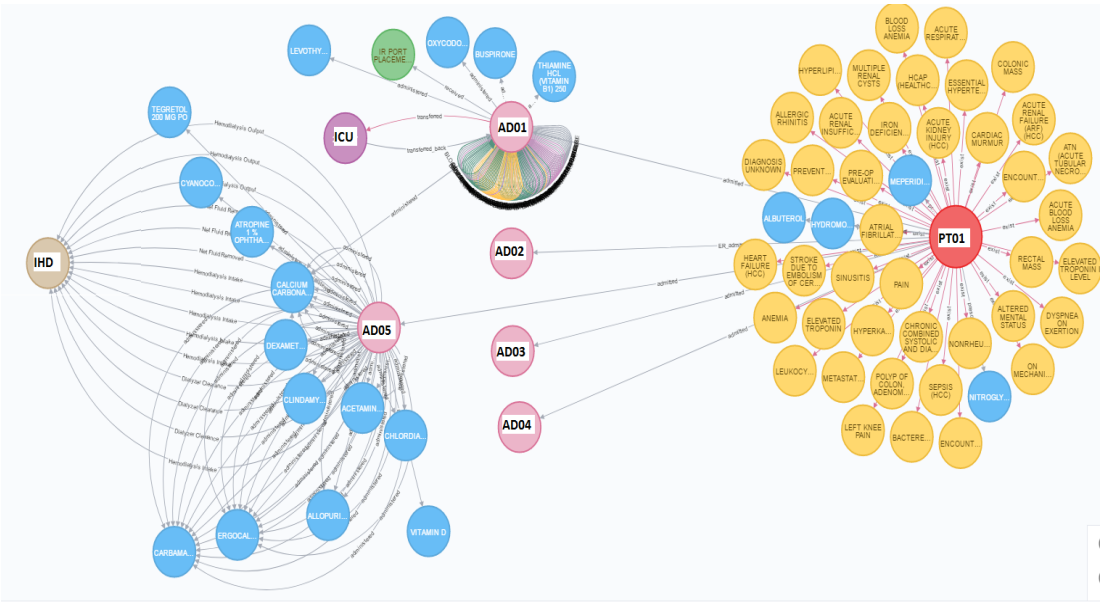

Fig. 17. Describes the entities and corresponding relationships. The colouring scheme further illustrates the interactive nature of linked data structures (Red: Patient, Yellow: Diagnosis, Blue: Medication, Pink: Hospital admission, Green: Procedure, Brown: Dialysis, and Purple: Intensive Care Unit). Patient PT01 has 5 hospital admissions (pink) - (AD01 through AD05), already has a number of pre-existing diagnoses (yellow), and 4 prescribed medications (blue on right). During hospital admission AD01, PT01 appears to be severely ill, is therefore transferred to the ICU, and had many repeated vitals measurements. During hospital admission AD05 (pink), PT01 developed acute kidney injury, received number of medications (blue left), and received intermittent hemodilaysis (IHD). Hospital admissions (AD02, AD03, and AD04) can be further discovered by exploding the respective nodes to learn about clinical events and interventions.

Table 4

Sample Graph Database Query

N.MRN	LOS	Dialysis_free	S.SURG_NAME
	133	67	"T 10 CORPECTOMY/ T4 LAMINECTOMY/ T5-L2 FUSION"
	133	51	"BONE MARROW BIOPSY"
	84	46	"VIDEO ASSISTED THORACOSCOPY - LOWER LOBECTOMY,
	64	44	"COLOSTOMY TAKEDOWN/REVERSAL"
	58	33	"OPEN CHOLECYSTECTOMY"
	36	31	"EXPLORATORY LAPAROTOMY"
	82	31	"ARTHROPLASTY REVISION HIP"
	60	30	"EGD"
	38	30	"EXPLORATORY LAPAROTOMY"
	42	29	"HARVEST VEIN OF LOWER EXTREMITIES"

Note. Demonstrates property graph model's data presentation capability. In addition to data visualization, a query results can be tabular. The table shows is a sample in descending order length of stay (LOS), free-dialysis days, and medical or surgical procedure performed - patients' identifiers are masked.

The inherent feature of drilling down in property graph models is demonstrated in Fig. 18. Two AKI patients share common diagnoses, AKI and acute respiratory failure, admitted twice to UAMS hospital, developed AKI and received dialysis (IHD) during one hospital admission. There are several shared clinical events between the two patients, but for visualization purposes, other nodes and relationships are rolled up. To learn more about a node or relationship, double clicking on the specific structure explodes and displays interrelated nodes and relationships.

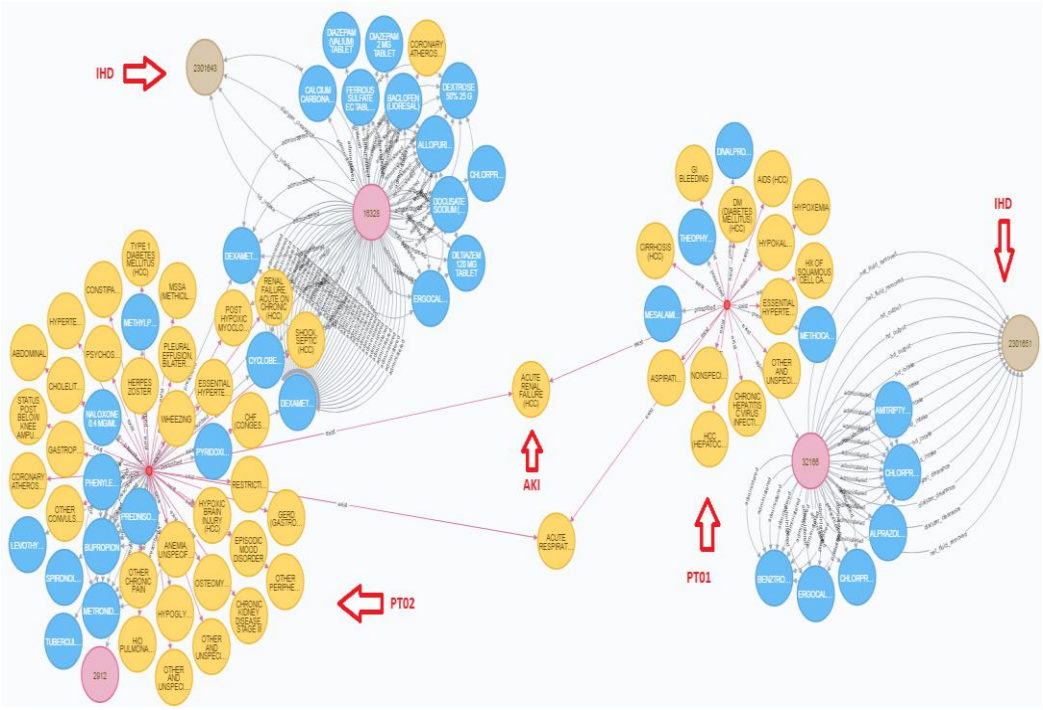

Fig. 18. Highlights the data exploratory feature of property graph model of AKI disease. Two patients PT01 and PT02 each had a hospital admission (pink coloured nodes) and during which each underwent hemodialysis (IHD). Both patients share at least two diagnoses, AKI and acute respiratory failure. Each hospital encounter (pink nodes) shows hospital length of stay (LOS) for that admission in minutes and can be converted to days by dividing by 1440. The drill down capability of a graph model are demonstrated by clicking on a node to view its dependencies and associated relationships.

B. OUTCOMES RISK ESTIMATION

In preparation for designing and implementing the outcome risk estimation tool, AKI patients' data are analysed and categorized, Table 1, Fig. 19, Fig 20, and Table 5, in the context of the three outcomes of interest: Alive and Dialysis Free, Alive and Dialysis Dependent, and Dead. Out of the 798 AKI patients, there were 243 patients that fully recovered (alive and dialysis free), 59 patients that recovered from AKI still need ongoing dialysis (alive and dialysis dependent). Patients on CRRT had worse 'DEAD' outcome (0.75 vs. 0.25) compared to patients who received IHD. The key clinical variables contributors to 'DEAD' outcome were: GCS < 15, sepsis, and SBP >100. The key clinical variables contributors to 'ALIVE AND DIALYSIS FREE' outcome were: RR < 22, GCS < 15, and CRRT. The key clinical variables contributors

to ‘ALIVE AND DIALYSIS DEPENDENT’ outcome were: SBP > 100, RR < 22, and IHD, Table 5. In clinical practice, sepsis may contribute up to 30% of acute kidney injury [73], and based on the data, it may be the single most individual clinical variable that leads to AKI. Although the data show that other clinical variables contribute more to ominous outcomes, which may be contributed by the presence of sepsis. Sepsis causes vasodilation which leads to decreased blood flow to the kidneys, increases respiratory rate, increases systolic blood pressure, and altered mentation. Respiratory rate and systolic blood pressure are also components of qSOFA (SBP <=100, RR >= 22, GCS < 15). Therefore, the combined effects of qSOFA components point to sepsis as contributor to poor outcomes. On the other hand, the same qSOFA clinical components (SBP > 100, RR<22, GCS=15) are key contributors to favourable outcomes - in descending order, SBP > 100 (0.80), RR (0.69), GCS (0.45). An aim of dialysis is fluid removal, and based on the limited amount of available fluid removal data, mortality appears lower in those who had fluid removed, but no significant difference between less or greater than one litre.

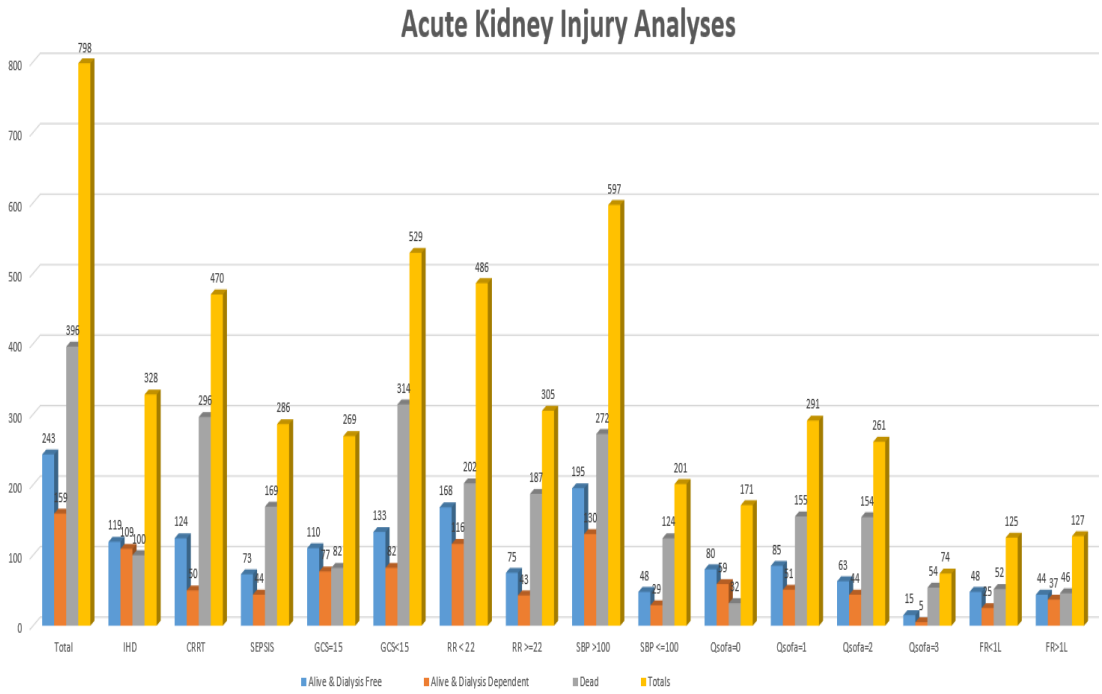

Fig. 19. Analysis of AKI raw data shows in respect to the three outcomes of interest patients with GCS (Glasgow Coma Score) have the highest mortality outcome followed by descending order, CRRT, SBP >100, and RR <22. However, it appears that patients with SBP>100 have favourable, alive and dialysis free, outcome, followed by those with those with RR <22. However, these numbers are skewed as they are reflected in the context of total patients for each of the clinical variables, e.g. 597 patients have SBP>100.

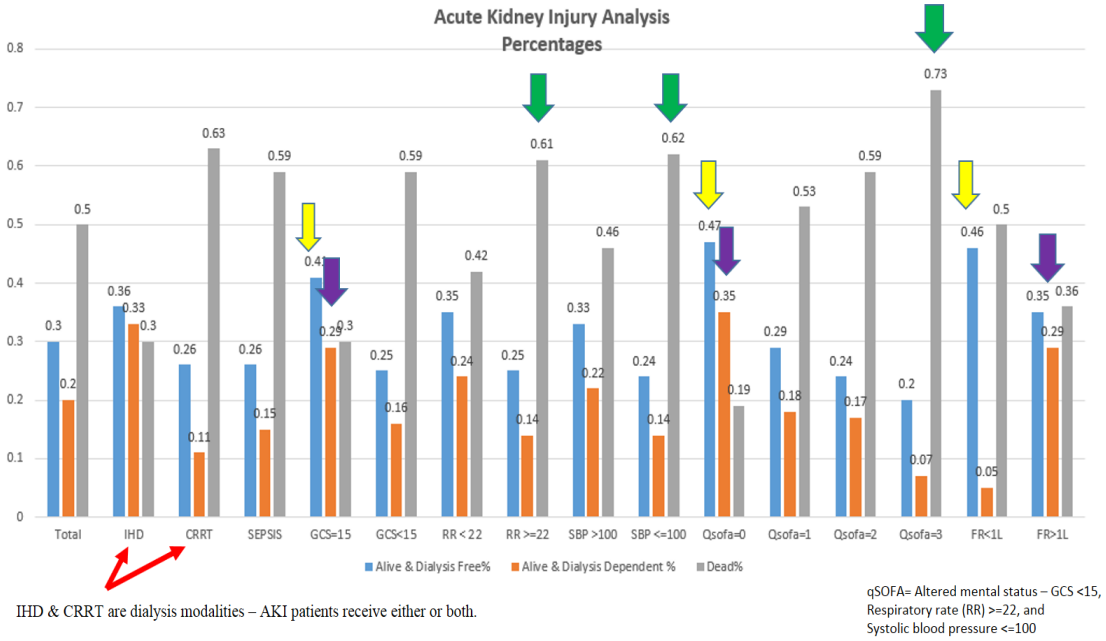

Fig. 20. The diagram shows AKI patients' outcomes. Based on the AKI cohort (n=798), qSOFA= 3 has the highest percentage of mortality (green arrow), followed by SBP <=100, RR >= 22, and GCS <15, which are the components of qSOFA. Clinically, that makes sense as patients with qSOFA score equals 3 are very sick and have unfavourable outcome. Conversely, patients with qSOFA equals to 0 (yellow arrow) have favourable outcome, alive and dialysis free, followed by FR (fluid removal) < 1L during initial dialysis and GCS=15. Overall, patients with qSOFA= 0 have the best outcomes, about 19% mortality, about 46% alive and dialysis free, and about 35% alive and dialysis dependent.

Next, the respective probability of each outcome for each clinical variable is calculated in preparation to calculate individual and/ or combined probabilities using Bayes' theorem. Probabilities in Table 5 constitute prior probability, and the calculated probabilities using the tool constitute posterior probabilities. And as more data flow into the AKI database, both prior (the evidence) and the posterior (outcome) probabilities are updated. It is worth noting that posterior probabilities often depend on more than one piece of evidence. For example, the probability of 'Dead' outcome of AKI patients varies and depends on whether a patient received IHD or CRRT dialysis, RR greater or less than 22, or GCS less than or equal to 15. The application of the implemented Bayesian Outcome Risk tool is clinically important but not warranted. Clinician can use the tool to estimate a patient's outcome risks based on available clinical variables stated in this study.

Table 5
Calculated Clinical Variables Probabilities

Clinical Variable	PER_ Free	PER_ Dep	PER_Dead	PER_NOT_ Free	PER_NOT_ Dep	PER_NOT_ Dead
Total	0.3	0.2	0.5	0.7	0.8	0.5
IHD	0.49	0.69	0.25	0.38	0.34	0.57
CRRT	0.51	0.31	0.75	0.62	0.66	0.43
SEPSIS	0.3	0.28	0.43	0.38	0.38	0.29
GCS=15	0.45	0.48	0.21	0.29	0.3	0.47
GCS<15	0.55	0.52	0.79	0.71	0.7	0.53
RR < 22	0.69	0.73	0.51	0.57	0.58	0.71
RR >=22	0.31	0.27	0.47	0.41	0.41	0.29
SBP >100	0.8	0.82	0.69	0.72	0.73	0.81
SBP <=100	0.2	0.18	0.31	0.28	0.27	0.19
Qsofa=0	0.33	0.37	0.08	0.16	0.18	0.35
Qsofa=1	0.35	0.32	0.39	0.37	0.38	0.34
Qsofa=2	0.26	0.28	0.39	0.36	0.34	0.27
Qsofa=3	0.06	0.03	0.14	0.11	0.11	0.05
FR<1L	0.2	0.16	0.13	0.14	0.16	0.18
FR>1L	0.18	0.23	0.12	0.15	0.14	0.2

Note. Provides probabilities for each of the clinical variable in the context of outcomes of interest in preparation for use in Bayes' conditional probability model.

The user interface tool, Fig. 14, displays a set of boxed clinical variables to select from for estimating outcomes. Each box contains reference to same clinical variable but with different scores. Therefore, only one selection is possible from each box, except for the box that contains the two types of dialysis, IHD and CRRT. An AKI patient may receive either dialysis modality or both – CRRT is usually given to severely ill patients in the ICU.

The conditional probabilities calculated by the tool for different clinical variables are compared with results from Fig. 19 and Fig. 20, and were validated with domain expert to ascertain their credibility. Fig. 21.1 through Fig.21.25 demonstrate different outcome risks probabilities for various clinical feature combinations. There are seven clinical features highlighting joint probabilities for AKI outcomes of interest. The probability calculations should be interpreted as follows: Aside from being 'Dead', the probabilities for 'alive and dialysis free' and 'alive and dialysis dependent' add to 1.

Clinical Variable	IHD √	SEPSIS	GCS	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.52		0.48			0.30	

Fig 21.1 Aside from the 0.30 dead probability, the remaining 0.70 probability is divided between the other two outcomes. The resulting probabilities are consistent with findings described by the AKI data analyses.

Clinical Variable	IHD √	SEPSIS √	GCS	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.53		0.47			0.39	

Fig 21.2 the inclusion of SEPSIS increases probability of mortality, and those that recover have slight increase in being dialysis dependent.

Clinical Variable	IHD √	SEPSIS √	GCS = 15	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.52		0.48			0.23	

Fig 21.3 GCS of 15 indicates patients are fully conscious, and therefore have decrease in mortality. The combination of GCS= 15. Sepsis normally increases ominous outcome but, in this case, the patient has GCS =15, indicating that patient is recovering from sepsis. However, there is increased probability of recovering and being dialysis dependent.

Clinical Variable	IHD √	SEPSIS √	GCS <15	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.55		0.45			0.49	

Fig 21.4 the presence of sepsis and GCS<15 indicate increased mortality, and if a patient survives, there is a split between being dialysis free and dialysis dependent.

Clinical Variable	IHD √	SEPSIS √	GCS	RR	SBP	QSOFA = 2	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.51		0.49			0.48	

Fig 21.5 qSOFA of 2 or 3 is associated with unfavourable outcomes. There is almost a split among the three outcomes.

Clinical Variable	IHD √	SEPSIS √	GCS	RR	SBP	QSOFA = 3	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.70		0.30			0.65	

Fig 21.6 QSOFA of 2 or 3 is associated with higher mortality. a qSOFA= 3 and sepsis indicate increase probability of mortality

Clinical Variable	IHD √	SEPSIS	GCS	RR	SBP	QSOFA	FR >1 L
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.45		0.55			0.21	

Fig 21.7 fluid removal appears to improve outcomes, at least decreased probability of mortality, though those surviving may have increased probability of being dialysis dependent. The finding confirms clinical expert finding that FR has positive impact on patient 'DEAD' outcome.

Clinical Variable	IHD √	SEPSIS	GCS	RR	SBP	QSOFA	FR <1 L
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.57			0.43			0.24

Fig 21.8 FR significantly decreases probability of mortality, but there is a split between being alive and dialysis free and alive and dialysis dependent.

Clinical Variable	CRRT √	SEPSIS	GCS	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.71			0.29			0.64

Fig 21.9 patients receiving CRRT are usually very ill and tend to have higher mortality rate than those on IHD. The probability is consistent with data analysis performed on the data set.

Clinical Variable	CRRT √	SEPSIS √	GCS	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.73			0.27			0.72

Fig 21.10 sepsis, when combined with CCRT, increases probability of mortality. The probability is consistent with data analysis on the data.

Clinical Variable	CRRT √	SEPSIS	GCS	RR	SBP	QSOFA	FR < 1L
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.76			0.24			0.56

Fig 21.11 CRRT patients are usually severely sick, treated in the ICU. FR improves 'DEAD' outcome but I still fairly high.

Clinical Variable	CRRT √	SEPSIS	GCS	RR	SBP	QSOFA	FR >1L
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.66			0.34			0.51

Fig 21.12 Again, fluid removal has positive impact on mortality outcome, but result show that fluid removal of < 1L tends to lead being alive and dialysis free slightly higher than > 1L.

Clinical Variable	CRRT √	SEPSIS	GCS	RR	SBP	QSOFA =0	FR
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.69			0.31			0.29

Fig 21.13 qsofa = 0, meaning patients have normal RR, SBP, and mentally alert, has positive impact across all outcomes. This can be indicative that a patient may be ready to be moved out of ICU.

Clinical Variable	CRRT √	SEPSIS	GCS	RR	SBP	QSOFA =2	FR
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.70			0.30			0.72

Fig 21.14 CRRT combined with qosfa =2 or 3 increases mortality. Patients have altered mental status, SBP < 100, and RR > 22.

Clinical Variable	CRRT √	SEPSIS	GCS	RR	SBP >100	QSOFA	FR
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent		Dead	
Probabilities	0.73			0.27		0.74	

Fig 21.15 SBP > 100, a component of qSOFA =2 or 3, indicated unfavourable mortality outcome.

Clinical Variable	CRRT √	SEPSIS √	GCS <15	RR >22	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.76		0.24			0.86	

Fig 21.16 the combination of the clinical variables produces very unfavourable mortality probability. CRRT and sepsis alone lead to ominous outcome, and RR > 22 indicate a patient may also have a qSOFA =2 or 3.

Clinical Variable	CRRT √	SEPSIS	GCS	RR >=22	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent			Dead	
Probabilities	0.74		0.26			0.74	

Fig 21.17 CRRT with RR >= 22 produces unfavourable mortality outcome. RR >22, a component of qSOFA, indicates a patient may have infection or possibly sepsis.

Clinical Variable	IHD & CRRT √	SEPSIS	GCS	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.64			0.36			0.43

Fig 21.18 patient receiving both IHD & CRRT shows split in outcomes, though mortality remains fairly high. A patient may have been in ICU (CRRT) bit then moved to ward (IHD).

Clinical Variable	IHD & CRRT √	SEPSIS	GCS =15	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free			Alive & Dialysis Dependent			Dead
Probabilities	0.62			0.38			0.25

Fig 21.19 patients receiving IHD & CRRT and GCS = 15 have good outcomes.

Clinical Variable	IHD & CRRT √	SEPSIS	GCS <15	RR	SBP	QSOFA	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent				Dead
Probabilities	0.65		0.35				0.53

Fig 21.20 patients receiving IHD & CRRT and GCS <15 increases mortality and split between probability of being alive and dialysis free and alive and dialysis dependent.

Clinical Variable	IHD & CRRT √	SEPSIS	GCS	RR	SBP	QSOFA =2	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent				Dead
Probabilities	0.62		0.38				0.52

Fig 21.21 patients receiving IHD & CRRT and qsofa =2 or 3 increase risk of mortality.

Clinical Variable	IHD & CRRT √	SEPSIS	GCS	RR	SBP	QSOFA =2	FR <1 L
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent				Dead
Probabilities	0.67		0.33				0.44

Fig.21.22 FR has positive impact on outcomes but qSOFA > 0 indicative of disruption in hemodynamic status (possibly RR > 22 or SBP < 100).

Clinical Variable	IHD & CRRT √	SEPSIS	GCS	RR	SBP	QSOFA =0	FR <1 L
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent				Dead
Probabilities	0.66		0.34				0.11

Fig 21.23 qSOFA= 0 and fluid removal significantly improves outcomes.

Clinical Variable	IHD & CRRT √	SEPSIS	GCS	RR	SBP	QSOFA =0	FR >1 L
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent				Dead
Probabilities	0.55		0.45				0.09

Fig 21.24 qsofa =0 and fluid removal > 1L greatly improves outcomes. Mortality is lowest among all other

Clinical Variable	IHD & CRRT √	SEPSIS	GCS	RR <22	SBP >100	QSOFA	FR
Outcomes	Alive & Dialysis Free		Alive & Dialysis Dependent				Dead
Probabilities	0.65		0.35				0.47

Fig 21.25 patients receiving IHD & CRRT, RR < 22 and SBP > 100 produces split among outcomes.

Fig 21 describes outcome probabilities for a number of clinical variables combinations using the Bayesian Outcome Risk tool. Note: for each of the probabilities above, the 'Alive & Dialysis Free' and 'Alive & Dialysis Dependent' outcomes are calculated using alive AKI patients after excluding the 'Dead' outcome.

C. NON-AKI TO AKI TRANSITION PROBABILITIES

In addition to outcome risk probabilities using Bayes' theorem, there is also deep interest in estimating the probability of a hospitalized, non-AKI patient developing AKI, and thus requiring dialysis. A transition probability from one health state to another is calculated by raising a transition matrix to the power *t*, a Markov Chain fact. In other words, given a set of initial vector probabilities, say (1 0), where 1 is non-AKI state and 0 is AKI state, we can calculate the probability of moving to the AKI state at a future step 2, say 24 hours, by raising the transition matrix to the power of 2, and the transition probability at 48 hours is calculated by raising the transition matrix to the power of 3, Fig. 22. Also, Fig 23 shows transition probabilities for the same clinical variables without the initial vectors.

In reality, however, patients may present with more than just one clinical variable, and we need to estimate transition probabilities when two or more variables impact outcomes. Fig. 23 shows examples of the interplay of at least two clinical variables in estimating transition probabilities.

Initial Vector: ***GCS <15***	**STEP 2**	**STEP 3**	**STEP 4**
$(1 \quad 0) * \begin{pmatrix} 0.38 & 0.62 \\ 0.55 & 0.45 \end{pmatrix}$	$(0.49 \quad 0.51)$	$(0.47 \quad 0.53)$	$(0.47 \quad 0.53)$

Initial Vector: ***SEPSIS***	**STEP 2**	**STEP 3**	**STEP 4**
$(1 \quad 0) * \begin{pmatrix} 0.07 & 0.93 \\ 0.65 & 0.35 \end{pmatrix}$	$(0.60 \quad 0.40)$	$(0.30 \quad 0.70)$	$(0.48 \quad 0.52)$

Initial Vector: ***SBP <= 100***	**STEP 2**	**STEP 3**	**STEP 4**
$(1 \quad 0) * \begin{pmatrix} 0.20 & 0.80 \\ 0.45 & 0.55 \end{pmatrix}$	$(0.40 \quad 0.60)$	$(0.35 \quad 0.65)$	$(0.36 \quad 0.64)$

Initial Vector: ***RR >= 22***	**STEP 2**	**STEP 3**	**STEP 4**
$(1 \quad 0) * \begin{pmatrix} 0.16 & 0.84 \\ 0.40 & 0.60 \end{pmatrix}$	$(0.36 \quad 0.64)$	$(0.31 \quad 0.69)$	$(0.32 \quad 0.68)$

Fig. 22. Given an initial vector space of a patient is in Non-AKI state at time *t0*, Markov Chain transition probabilities are calculated at times *t1*, *t2*, and *t3* for each of the clinical variables.

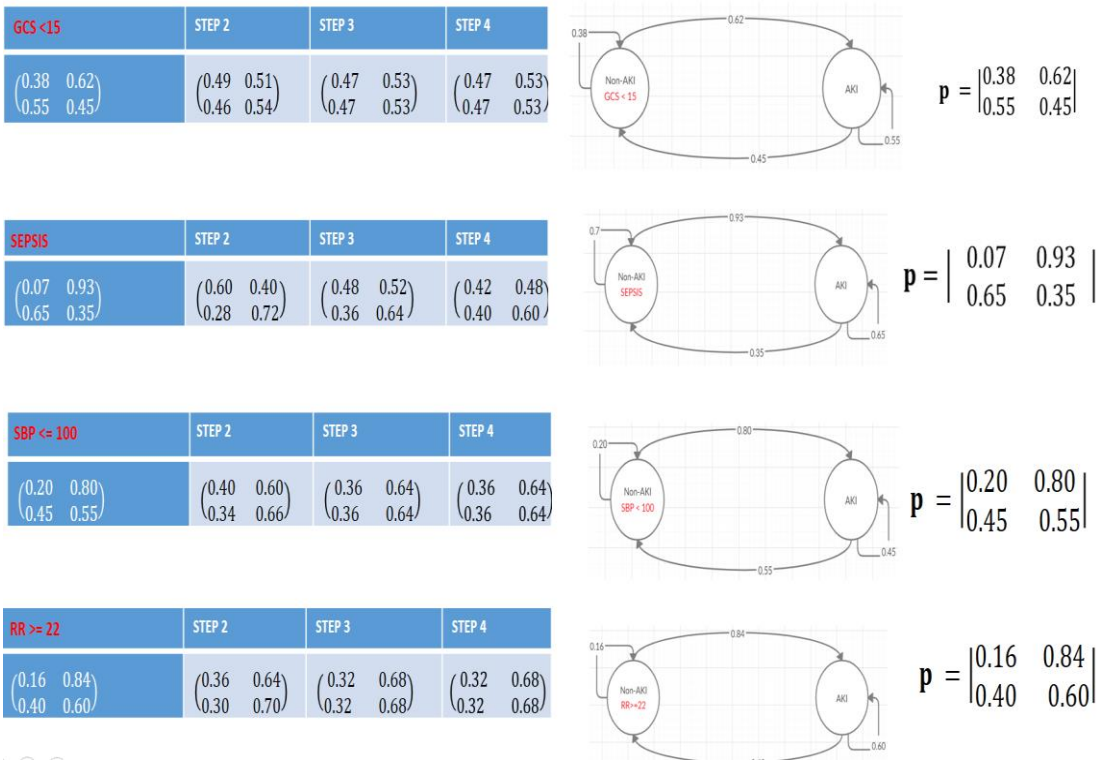

Fig. 23. The same clinical variables as in Fig 22 but without the initial vectors. Note that the transition probabilities at *t1*, *t2*, and *t3* are the same as in Fig 22, which are due the initial vector values (1 0).

The initial transition probabilities from a non-AKI to an AKI state are quite high, Fig 23. As expected, the indicated clinical variables are the ones that have the potential to contribute most to the development of AKI, Fig. 19 and Fig. 20. Interestingly, transition probabilities to an AKI state may fluctuate over future states but eventually tend to stabilize, a Markov's long run distribution property. The decrease in the transition probability over time can be justified clinically. For example, sepsis, a critical medical condition that requires immediate medical intervention, transition probability from *t0*=0.93 to *t1*=0.40 may indicate that at *t1* a patient may have either received appropriate treatments, e.g., antibiotics and dialysis, or the outcome is no longer favourable.

In many clinical cases, patients in critical care settings may present with multitude of signs and symptoms, and may undergo several and frequent laboratory measurements. Assessment of transition probabilities could involve two or more clinical variables. Fig. 24 shows different transition probabilities of clinical variables occurring simultaneously. Per Markov chain, the transition probabilities of combined probabilities decrease by some magnitude at every step but stabilize at some future step. However, the combined transition probabilities remain slightly higher compared to individual probabilities, Fig. 23.

Initial Vector GCS <15 & SEPSIS	STEP 2	STEP 3	STEP 4
$\begin{pmatrix} 0.54 & 0.46 \\ 0.44 & 0.56 \end{pmatrix}$	$\begin{pmatrix} 0.49 & 0.51 \\ 0.48 & 0.52 \end{pmatrix}$	$\begin{pmatrix} 0.49 & 0.51 \\ 0.49 & 0.51 \end{pmatrix}$	$\begin{pmatrix} 0.49 & 0.51 \\ 0.49 & 0.51 \end{pmatrix}$
Initial Vector **SEPSIS & RR>=22**	**STEP 2**	**STEP 3**	**STEP 4**
$\begin{pmatrix} 0.38 & 0.62 \\ 0.24 & 0.76 \end{pmatrix}$	$\begin{pmatrix} 0.29 & 0.71 \\ 0.27 & 73 \end{pmatrix}$	$\begin{pmatrix} 0.28 & 0.72 \\ 0.28 & 0.72 \end{pmatrix}$	$\begin{pmatrix} 0.28 & 0.72 \\ 0.28 & 0.72 \end{pmatrix}$

Fig. 24. Given an initial vector space of a patient is in Non-AKI state at time *t0*, Markov Chain transition probabilities are calculated at times *t1*, *t2*, and *t3* for combinations of clinical variables.

CHAPTER VI DISCUSSION

The adoption of electronic medical records and subsequent data deluge motivated clinical research communities to leverage analytic tools to discover knowledge. The continued and exponential growth of clinical data surpassed the unequal development of new methods to deliver the right data at the right time to conduct meaningful research. Clinical data warehouses have been the mainstay of storing and disseminating data for research; however, the one-size fit all relational backbone of present day data warehouses are not capable of providing the agility and rigor expected of clinical research. The informatics aspects of this study are timely as there is deep interest among nephrology clinical investigators to understand better the etiology of, and predictors of outcomes of acute kidney injury (AKI) patients. The lifecycle of extracting data form institutional data warehouse to analyse is often cyclical, repetitive, and time-consuming.

The aim of the study is to design a horizontally scalable database model of acute kidney injury clinical facts that resembles AKI patients' cycling through a hospital setting. The model uses graph methods to represent clinical knowledge and builds on successes in other domains such as social networks and bioinformatics. It aims to create a 360-view of AKI patients' recycling through a healthcare system by building meaningful and connected relationships between entities.

The results show that a property graph model provides the right platform to not only conduct clinical research but also serves as an excellent decision-support environment as well. The visualizations and exploration aspects of the AKI graph model are data-driven and clinicians and researchers alike can uncover knowledge that may not be

discernible otherwise. A clinician need not define a question and query; rather, he can visualize a patient, view similarities with other patients, and compile and make informed decisions about a patient's management plan.

While previous research has focused on using graph models in social networks and genomic studies, the results of this study confirm that graph models can be excellent for representing clinical events. However, the design and implementation of a clinical graph model require participation of a clinical informatician knowledgeable in clinical workflows and familiarity with knowledge management techniques.

The results of the study provide a new insight into the design and implementation of future clinical research data warehouses. Instead of implementing a one-size-fit all solution that stores all sorts of data from an EMR, it is more feasible to provide on-demand research model that addresses specific research needs that provides information-ready and actionable information. The model aligns with the concept of a data lake, where a data lake represents a specific clinical domain, and can coalesce with other data lakes – clinical domains- if a relationship exists between entities (nodes).

While a key aim of the study is to leverage the huge amount of acute kidney injury clinical data to learn about potential patients' outcomes, the AKI graph model proved useful in envisioning predictive models such as Bayesian inference and Markov Chain to estimate outcome probabilities. By slicing and dicing the AKI data, the study succeeded in categorizing outcomes in the context of several clinical variables, which served as the building blocks for the Bayesian outcome risk tool and Markov Chain state transitions. The conditional probabilities estimated by the Bayesian inference

correlate well within the context of the clinical variables, e.g. sepsis. Typically, in clinical practice, there are more than one clinical variable that could impact outcomes, and the Bayesian model is able to, as the results show, to take into account as many clinical variables as possible in estimating outcomes.

Estimating a random non-AKI patient probability of developing AKI – requiring dialysis- in 48 hours is tricky. Some of the clinical variables discussed earlier can be sensitive in influencing AKI patients' outcomes (alive and dialysis free, alive and dialysis dependent, or dead), they are not specific to acute kidney injury. The clinical variables such RR > 22 or SBP < 100 are common in many other medical conditions. The implementation demonstrated earlier of Markov Chain is conceptual, and the resulting transition probabilities are experimental. The Markov model, however, requires fine-tuning initial state probabilities and further testing.

The results strongly support the potential of using graph models to represent clinical data to address limitations of relational models. The AKI data represented in a graph model contributes to clearer understanding of AKI. The structural schema of a graph model motivated moving compute time to the source; that is, the implementation of analytic solutions, Bayesian inference and conceptualization of Markov Chain, to be part of the model structure. While the AKI data representation in the graph database are concordant with AKI data in the institutional data warehouse, the results of both Bayesian and Markov probabilities, while seem reasonable, require further validation and testing. And, as more acute kidney injury data become available, both the Bayesian and Markov Chain models' respective probabilities are updated.

Markov models are useful when a decision problem involves risk that is continuous over time, and when important events may happen more than once [70]. The study proposed a general model of estimating transition probabilities from a non- AKI state to AKI state using discrete clinical variables values. During a hospital stay, clinical variables such as blood pressure and respiratory rate are measured frequently to assess health status, and are recorded in electronic medical record system. Estimating transition probabilities at a point of time provides only a snapshot but is not strongly indicative of a future state transition. However, a transition probability calculated based on a clinical variable trajectory over a period of time, using continuous variable, would be much stronger, Fig. 25.

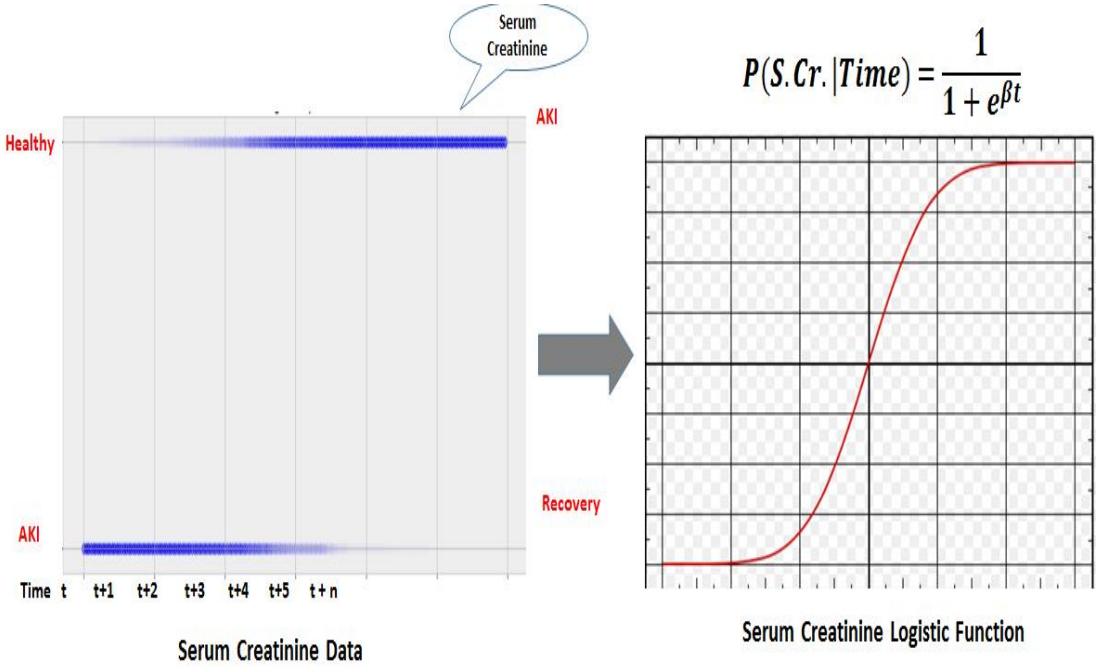

Fig. 25 Describes trajectory of serum creatinine changes from a healthy, non-AKI, state, to AKI state. Since there is no specific cut off serum creatinine value that determines transitioning from one state to another, the probability can be best represented with a logistic function using continuous variable.

Another factor that plays a significant role in determining transition states in Markov is initial probability of a state. In calculating initial probabilities for few clinical variables, Fig. 15, the odds ratio for each clinical variable was determined from a

contingency table, and the results show, Table 6, that the odds ratio for 'sepsis' is quite high and is indicative of a strong association with AKI. However, the z statistic for 'sepsis' is very high - z score falls outside the CI of 19.4 and 26.0 – which indicate the pattern exhibited is too unusual to be a random chance. In this case, it is possible to reject the null hypothesis, that there is no association between sepsis and AKI, and figure out what might be causing the statistically significant spatial pattern. The same argument applies to the other clinical variables stated in Table 6. Again, estimating initial probabilities are critical in determining Markov transition states and, as far as this study is concerned, reviewing source data, e.g. sepsis, from which initial probabilities are calculated, needs evaluated.

Table 6
Clinical Variables' Statistical Significance

Clinical Variable	**Odds Ratio**	**95% CI (P < 0.0001)**	**Z Statistic**
Sepsis	23	19.4 - 26.0	41.8
GCS < 15	1.7	1.4 – 1.9	6.0
SBP <=100	4.2	3.6 – 5.0	17.8
RR >=22	5.3	4.6 – 6.1	22.7

Note. Statistical relevance of few clinical variables within the context of acute kidney injury.

CHAPTER VII CONCLUSION

This research aimed primarily to represent acute kidney injury (AKI) clinical knowledge in an effective and efficient model for streamlining conduct of clinical research. Secondarily, this research tried to answer key acute kidney injury patients' outcomes using Bayes' theorem, and to predict patients' likelihood of requiring dialysis in 48 hours in patients without the disease using Markov Chain state transition probabilities. Based on the acute kidney injury property graph implementation and visualization capability of query results, it can be concluded that the primary aim of the research has been met, and provides an improved solution to represent, query, and visualize clinical knowledge. And, based on the quantitative solution to answer AKI and non-AKI outcome probabilities, it can be concluded that the secondary aims are also met and both provide novel analytic approaches in the context of graph models. The results indicate that a property graph model has the potential to serve targeted clinical research needs. The integration of analytic tools with property graph models, moving compute time to the source, enhances functionality and shortens research lifecycle. The upfront investment in standing up clinically -oriented property graph model requires mastery of knowledge management methodologies and the clinical subject area.

Markov Chain provides a simple model to estimate transition states from a non-AKI to AKI state; however, estimating initial transition probabilities may profoundly affect the model's accuracy. A planned, future work includes investigating use of a Bayesian network model, Fig. 4, to estimate probability of transitioning from a non-AKI to AKI state and define initial probabilities based on AKI domain expert.

Within the context of the AKI property graph model, future work would attempt to apply graph theory's similarity algorithms to calculate (dis)similarities between two sets of data. One algorithm, Euclidean distance algorithm measures the straight-line distance between two points in n-dimensional space, and Overlap algorithm measures overlap between two sets, Fig 26. The Overlap algorithm could point out differences and similarities between two sets of patients sharing similar outcomes.

```
MATCH (p1:INPATIENT {MRN:      '})-[likes1:assign]->(DIAGNOSIS)
MATCH (p2:INPATIENT {MRN:      '})-[likes2:assign]->(DIAGNOSIS)
RETURN p1.MRN AS from,
    p2.MRN AS to,
    algo.similarity.euclideanDistance(collect(likes1.DX_DATE_R),
collect(likes2.DX_DATE_R)) AS similarity

from to    similarity
                39388072.96233722
```

A. Euclidean Distance Algorithm

```
MATCH (P:INPATIENT)-[:transferred]->(ICU)
WITH {item:id(ICU), categories: collect(id(P))} as userData
WITH collect(userData) as data
CALL algo.similarity.overlap.stream(data)
YIELD item1, item2, count1, count2, intersection, similarity
RETURN algo.asNode(item1).MRN AS from, algo.asNode(item2).MRN
AS to,
    count1, count2, intersection, similarity
ORDER BY similarity DESC;
```

from	to	count1	count2	intersection	similarity
		1	1	0	0.0
		1	1	0	0.0
		1	1	0	0.0
		1	1	0	0.0

B. Overlap Similarity Algorithm

Fig. 26. Highlight use of graph's theory similarity algorithms. A. Euclidean Distance Algorithm measures distance between two points in a graph, and B. Overlap Similarity Algorithm measures overlap between two data sets within a graph.

REFERENCES

1. Alocci, D., Mariethoz, J., Horlacher, O., Bolleman, J. T., Campbell, M. P., & Lisacek, F. (2015). Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search. *PLoS One, 10*(12), e0144578. doi: 10.1371/journal.pone.0144578

2. Arous, E. J., McDade, T. P., Smith, J. K., Ng, S. C., Sullivan, M. E., Zottola, R. J., . . . Tseng, J. F. (2014). Electronic medical record: research tool for pancreatic cancer? *J Surg Res, 187*(2), 466-470. doi: 10.1016/j.jss.2013.10.036

3. Bae, C. J., Griffith, S., Fan, Y., Dunphy, C., Thompson, N., Urchek, J., . . . Katzan, I. L. (2015). The Challenges of Data Quality Evaluation in a Joint Data Warehouse. *EGEMS (Wash DC), 3*(1), 1125. doi: 10.13063/2327-9214.1125

4. Balaur, I., Mazein, A., Saqi, M., Lysenko, A., Rawlings, C. J., & Auffray, C. (2017). Recon2Neo4j: applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics, 33*(7), 1096-1098. doi: 10.1093/bioinformatics/btw731

5. Barros, M., & Couto, F. M. (2016a). Knowledge Representation and Management: a Linked Data Perspective. *Yearb Med Inform*(1), 178-183. doi: 10.15265/IY-2016-022

6. Bauer, C. R., Ganslandt, T., Baum, B., Christoph, J., Engel, I., Löbe, M., . . . Sax, U. (2016). Integrated Data Repository Toolkit (IDRT). A Suite of Programs to Facilitate Health Analytics on Heterogeneous Medical Data. *Methods Inf Med, 55*(2), 125-135. doi: 10.3414/ME15-01-0082

7. Bottomly, D., McWeeney, S. K., & Wilmot, B. (2016). HitWalker2: visual analytics for precision medicine and beyond. *Bioinformatics, 32*(8), 1253-1255. doi: 10.1093/bioinformatics/btv739

8. Chen, Y. A., Tripathi, L. P., & Mizuguchi, K. (2016). An integrative data analysis platform for gene set analysis and knowledge discovery in a data warehouse framework. *Database (Oxford), 2016*. doi: 10.1093/database/baw009

9. Croset, S., Rupp, J., & Romacker, M. (2016). Flexible data integration and curation using a graph-based approach. *Bioinformatics, 32*(6), 918-925. doi: 10.1093/bioinformatics/btv644

10. Dai, X., Li, J., Liu, T., & Zhao, P. X. (2016). HRGRN: A Graph Search-Empowered Integrative Database of Arabidopsis Signaling Transduction, Metabolism and Gene Regulation Networks. *Plant Cell Physiol, 57*(1), e12. doi: 10.1093/pcp/pcv200

11. Danciu, I., Cowan, J. D., Basford, M., Wang, X., Saip, A., Osgood, S., . . . Harris, P. A. (2014). Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform, 52*, 28-35. doi: 10.1016/j.jbi.2014.02.003

12. Denney, M. J., Long, D. M., Armistead, M. G., Anderson, J. L., & Conway, B. N. (2016). Validating the extract, transform, load process used to populate a large clinical research database. *Int J Med Inform, 94*, 271-274. doi: 10.1016/j.ijmedinf.2016.07.009

13. Evans, R. S., Lloyd, J. F., & Pierce, L. A. (2012). Clinical use of an enterprise data warehouse. *AMIA Annu Symp Proc, 2012*, 189-198.

14. Gesicho, M. B., Babic, A., & Were, M. C. (2017). Critical Issues in Evaluating National-Level Health Data Warehouses in LMICs: Kenya Case Study. *Stud Health Technol Inform, 238*, 201-204.

15. Geva, A., Gronsbell, J. L., Cai, T., Murphy, S. N., Lyons, J. C., Heinz, M. M., . . . Pediatric Pulmonary Hypertension Network and National Heart, L. n., and Blood Institute Pediatric Pulmonary Vascular Disease Outcomes Bioinformatics Clinical Coordinating Center Investigators. (2017a). A Computable Phenotype Improves Cohort Ascertainment in a Pediatric Pulmonary Hypertension Registry. *J Pediatr, 188*, 224-231.e225. doi: 10.1016/j.jpeds.2017.05.037

16. Gkoutos, G. V., Schofield, P. N., & Hoehndorf, R. (2017). The anatomy of phenotype ontologies: principles, properties and applications. *Brief Bioinform*. doi: 10.1093/bib/bbx035

17. Godderis, L., Mylle, G., Coene, M., Verbeek, C., Viaene, B., Bulterys, S., & Schouteden, M. (2015). Data warehouse for detection of occupational diseases in OHS data. *Occup Med (Lond), 65*(8), 651-658. doi: 10.1093/occmed/kqv074

18. Guo, Z., Kashyap, S., Sonka, M., & Oguz, I. (2017). Machine learning in a graph framework for subcortical segmentation. *Proc SPIE Int Soc Opt Eng, 10133*. doi: 10.1117/12.2254874

19. Haarbrandt, B., Tute, E., & Marschollek, M. (2016). Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. *J Biomed Inform, 63*, 277-294. doi: 10.1016/j.jbi.2016.08.007

20. Haque, W., Urquhart, B., Berg, E., & Dhanoa, R. (2014). Using business intelligence to analyze and share health system infrastructure data in a rural health authority. *JMIR Med Inform, 2*(2), e16. doi: 10.2196/medinform.3590

21. He, T., & Chan, K. C. (2016). Evolutionary Graph Clustering for Protein Complex

Identification. *IEEE/ACM Trans Comput Biol Bioinform*. doi: 10.1109/TCBB.2016.2642107

22. Horton, I., Lin, Y., Reed, G., Wiepert, M., & Hart, S. (2017). Empowering Mayo Clinic Individualized Medicine with Genomic Data Warehousing. *J Pers Med, 7*(3). doi: 10.3390/jpm7030007

23. Househ, M., & Aldosari, B. (2017). The Hazards of Data Mining in Healthcare. *Stud Health Technol Inform, 238*, 80-83.

24. Hristovski, D., Kastrin, A., Dinevski, D., & Rindflesch, T. C. (2015). Constructing a Graph Database for Semantic Literature-Based Discovery. *Stud Health Technol Inform, 216*, 1094.

25. Hu, Y., Liu, H., Du, L., Wan, J., & Li, X. (2017). Serum Cystatin C Predicts AKI and the Prognosis of Patients in Coronary Care Unit: a Prospective, Observational Study. *Kidney Blood Press Res, 42*(6), 961-973. doi: 10.1159/000485341

26. Johnson, D., Connor, A. J., McKeever, S., Wang, Z., Deisboeck, T. S., Quaiser, T., & Shochat, E. (2014a). Semantically linking in silico cancer models. *Cancer Inform, 13*(Suppl 1), 133-143. doi: 10.4137/CIN.S13895

27. Kamal, J., Liu, J., Ostrander, M., Santangelo, J., Dyta, R., Rogers, P., & Mekhjian, H. S. (2010). Information warehouse - a comprehensive informatics platform for business, clinical, and research applications. *AMIA Annu Symp Proc, 2010*, 452-456.

28. Karami, M., Rahimi, A., & Shahmirzadi, A. H. (2017). Clinical Data Warehouse: An Effective Tool to Create Intelligence in Disease Management. *Health Care Manag (Frederick), 36*(4), 380-384. doi: 10.1097/HCM.0000000000000113

29. Karmen, C., Ganzinger, M., Kohl, C. D., Firnkorn, D., & Knaup-Gregori, P. (2014). A framework for integrating heterogeneous clinical data for a disease area into a central data warehouse. *Stud Health Technol Inform, 205*, 1060-1064.

30. Kennell, T., Dempsey, D. M., & Cimino, J. J. (2016). i3b3: Infobuttons for i2b2 as a Mechanism for Investigating the Information Needs of Clinical Researchers. *AMIA Annu Symp Proc, 2016*, 696-704.

31. Kimmel, L. A., Wilson, S., Walker, R. G., Singer, Y., & Cleland, H. (2017). Acute Kidney Injury: It's not just the 'big' burns. *Injury*. doi: 10.1016/j.injury.2017.11.016

32. Kocbek, S., & Kim, J. D. (2017). Exploring biomedical ontology mappings with graph theory methods. *PeerJ, 5*, e2990. doi: 10.7717/peerj.2990

33. Kocevar, G., Stamile, C., Hannoun, S., Cotton, F., Vukusic, S., Durand-Dubief, F., & Sappey-Marinier, D. (2016). Graph Theory-Based Brain Connectivity for Automatic Classification of Multiple Sclerosis Clinical Courses. *Front Neurosci, 10*, 478. doi: 10.3389/fnins.2016.00478

34. Kock-Schoppenhauer, A. K., Kamann, C., Ulrich, H., Duhm-Harbeck, P., & Ingenerf, J. (2017). Linked Data Applications Through Ontology Based Data Access in Clinical Research. *Stud Health Technol Inform, 235*, 131-135.

35. Kozaki, K., Yamagata, Y., Mizoguchi, R., Imai, T., & Ohe, K. (2017). Disease Compass- a navigation system for disease knowledge based on ontology and linked data techniques. *J Biomed Semantics, 8*(1), 22. doi: 10.1186/s13326-017-0132-2

36. Kunz, S. N., Zupancic, J. A. F., Rigdon, J., Phibbs, C. S., Lee, H. C., Gould, J. B., . . . Profit, J. (2017). Network analysis: a novel method for mapping neonatal acute transport patterns in California. *J Perinatol, 37*(6), 702-708. doi: 10.1038/jp.2017.20

37. Lasierra, N., Schweitzer, M., Gorfer, T., Toma, I., & Hoerbst, A. (2016). Building a Semantic Model to Enhance the User's Perceived Functionality of the EHR. *Stud Health Technol Inform, 228*, 137-141.

38. Lee, G., Lee, H. B., Jung, B. H., & Nam, H. (2017). mvp - an open-source preprocessor for cleaning duplicate records and missing values in mass spectrometry data. *FEBS Open Bio, 7*(7), 1051-1059. doi: 10.1002/2211-5463.12247

39. Lozano-Rubí, R., Pastor, X., & Lozano, E. (2014). OWLing Clinical Data Repositories With the Ontology Web Language. *JMIR Med Inform, 2*(2), e14. doi: 10.2196/medinform.3023

40. Lysenko, A., Roznovăţ, I. A., Saqi, M., Mazein, A., Rawlings, C. J., & Auffray, C. (2016). Representing and querying disease networks using graph databases. *BioData Min, 9*, 23. doi: 10.1186/s13040-016-0102-8

41. Madkour, M., Benhaddou, D., & Tao, C. (2016). Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain. *Comput Methods Programs Biomed, 128*, 52-68. doi: 10.1016/j.cmpb.2016.02.007

42. Maxwell, R. A., & Bell, C. M. (2017). Acute Kidney Injury in the Critically Ill. *Surg Clin North Am, 97*(6), 1399-1418. doi: 10.1016/j.suc.2017.07.004

43. Mironov, V., Seethappan, N., Blondé, W., Antezana, E., Splendiani, A., & Kuiper, M. (2012). Gauging triple stores with actual biological data. *BMC Bioinformatics, 13 Suppl 1*, S3. doi: 10.1186/1471-2105-13-S1-S3

44. Mughal, S., Moghul, I., Yu, J., Clark, T., Gregory, D. S., & Pontikos, N. (2017). Pheno4J: a gene to phenotype graph database. *Bioinformatics*. doi: 10.1093/bioinformatics/btx397

45. Mullen, J., Cockell, S. J., Woollard, P., & Wipat, A. (2016). An Integrated Data Driven Approach to Drug Repositioning Using Gene-Disease Associations. *PLoS One, 11*(5), e0155811. doi: 10.1371/journal.pone.0155811

46. Muslem, R., Caliskan, K., Akin, S., Sharma, K., Gilotra, N. A., Constantinescu, A. A., . . . Manintveld, O. C. (2017). Acute kidney injury and 1-year mortality after left ventricular assist device implantation. *J Heart Lung Transplant*. doi: 10.1016/j.healun.2017.11.005

47. Ningthoujam, S. S., Choudhury, M. D., Potsangbam, K. S., Chetia, P., Nahar, L., Sarker, S. D., . . . Das Talukdar, A. (2014). NoSQL data model for semi-automatic integration of ethnomedicinal plant data from multiple sources. *Phytochem Anal, 25*(6), 495-507. doi: 10.1002/pca.2520

48. Odgers, D. J., & Dumontier, M. (2015). Mining Electronic Health Records using Linked Data. *AMIA Jt Summits Transl Sci Proc, 2015*, 217-221.

49. Parmanto, B., Scotch, M., & Ahmad, S. (2005). A framework for designing a healthcare outcome data warehouse. *Perspect Health Inf Manag, 2*, 3.

50. Pokorny, J. J., Norman, A., Zanesco, A. P., Bauer-Wu, S., Sahdra, B. K., & Saron, C. D. (2017). Network Analysis for the Visualization and Analysis of Qualitative Data. *Psychol Methods*. doi: 10.1037/met0000129

51. Post, A. R., Kurc, T., Cholleti, S., Gao, J., Lin, X., Bornstein, W., . . . Saltz, J. H. (2013). The Analytic Information Warehouse (AIW): a platform for analytics using electronic health record data. *J Biomed Inform, 46*(3), 410-424. doi: 10.1016/j.jbi.2013.01.005

52. Richardson, K. L., Watson, R. S., & Hingorani, S. (2017). Quality of life following hospitalization-associated acute kidney injury in children. *J Nephrol*. doi: 10.1007/s40620-017-0450-6

53. Roelofs, E., Persoon, L., Nijsten, S., Wiessler, W., Dekker, A., & Lambin, P. (2013). Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiother Oncol, 108*(1), 174-179. doi: 10.1016/j.radonc.2012.09.019

54. Roski, J., Bo-Linn, G. W., & Andrews, T. A. (2014). Creating value in health care through big data: opportunities and policy implications. *Health Aff (Millwood),*

33(7), 1115-1122. doi: 10.1377/hlthaff.2014.0147

55. Shin, S. Y., Lyu, Y., Shin, Y., Choi, H. J., Park, J., Kim, W. S., & Lee, J. H. (2013). Experience of de-identification system development for clinical research in tertiary hospital. *Stud Health Technol Inform, 192*, 1044.

56. Spratt, S. E., Pereira, K., Granger, B. B., Batch, B. C., Phelan, M., Pencina, M., . . . Group, D. P. (2017). Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. *J Am Med Inform Assoc, 24*(e1), e121-e128. doi: 10.1093/jamia/ocw123

57. Summer, G., Kelder, T., Ono, K., Radonjic, M., Heymans, S., & Demchak, B. (2015). cyNeo4j: connecting Neo4j and Cytoscape. *Bioinformatics, 31*(23), 3868-3869. doi: 10.1093/bioinformatics/btv460

58. Sun, H., Depraetere, K., De Roo, J., Mels, G., De Vloed, B., Twagirumukiza, M., & Colaert, D. (2015a). Semantic processing of EHR data for clinical research. *J Biomed Inform, 58*, 247-259. doi: 10.1016/j.jbi.2015.10.009

59. Sun, H., Depraetere, K., De Roo, J., Mels, G., De Vloed, B., Twagirumukiza, M., & Colaert, D. (2015b). Semantic processing of EHR data for clinical research. *J Biomed Inform, 58*, 247-259. doi: 10.1016/j.jbi.2015.10.009

60. Turley, C. B., Obeid, J., Larsen, R., Fryar, K. M., Lenert, L., Bjorn, A., . . . Sanderson, I. (2016). Leveraging a Statewide Clinical Data Warehouse to Expand Boundaries of the Learning Health System. *EGEMS (Wash DC), 4*(1), 1245. doi: 10.13063/2327-9214.1245

61. Wiesenauer, M., Johner, C., & Röhrig, R. (2012). Secondary use of clinical data in healthcare providers - an overview on research, regulatory and ethical requirements. *Stud Health Technol Inform, 180*, 614-618.

62. Wunsch, G., da Costa, C. A., & Righi, R. R. (2017). A Semantic-Based Model for Triage Patients in Emergency Departments. *J Med Syst, 41*(4), 65. doi: 10.1007/s10916-017-0710-y

63. Yoo, S., Hwang, H., & Jheon, S. (2016). Hospital information systems: experience at the fully digitized Seoul National University Bundang Hospital. *J Thorac Dis, 8*(Suppl 8), S637-641. doi: 10.21037/jtd.2016.08.44

64. Yoo, S., Kim, S., Lee, K. H., Jeong, C. W., Youn, S. W., Park, K. U., . . . Hwang, H. (2014). Electronically implemented clinical indicators based on a data warehouse in a tertiary hospital: its clinical benefit and effectiveness. *Int J Med Inform, 83*(7), 507-516. doi: 10.1016/j.ijmedinf.2014.04.001

65. Yoon, B. H., Kim, S. K., & Kim, S. Y. (2017). Use of Graph Database for the Integration of Heterogeneous Biological Data. *Genomics Inform, 15*(1), 19-27. doi: 10.5808/GI.2017.15.1.19

66. Zolhavarieh, S., Parry, D., & Bai, Q. (2017). Issues Associated With the Use of Semantic Web Technology in Knowledge Acquisition for Clinical Decision Support Systems: Systematic Review of the Literature. *JMIR Med Inform, 5*(3), e18. doi: 10.2196/medinform.6169

67. Zorrilla-Vaca, A., Ziai, W., Connolly, E. S., Geocadin, R., Thompson, R., & Rivera-Lara, L. (2017). Acute Kidney Injury Following Acute Ischemic Stroke and Intracerebral Hemorrhage: A Meta-Analysis of Prevalence Rate and Mortality Risk. *Cerebrovasc Dis, 45*(1-2), 1-9. doi: 10.1159/000479338

68. Olariu, E., Cadwell, K. K., Hancock, E., Trueman, D., & Chevrou-Severac, H. (2017). Current recommendations on the estimation of transition probabilities in Markov cohort models for use in health care decision-making: a targeted literature review. *Clinicoecon Outcomes Res, 9*, 537-546. doi: 10.2147/CEOR.S135445

69. Luo, L., Zhang, F., Zhang, W., Sun, L., Li, C., Huang, D., . . . Wang, B. (2017). Markov Chain-Based Acute Effect Estimation of Air Pollution on Elder Asthma Hospitalization. *J Healthc Eng, 2017*, 2463065. doi: 10.1155/2017/2463065

70. Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making: a practical guide. *Med Decis Making, 13*(4), 322-338. doi: 10.1177/0272989X9301300409

71. Hsu, R. K., McCulloch, C. E., Dudley, R. A., Lo, L. J., & Hsu, C. Y. (2013). Temporal changes in incidence of dialysis-requiring AKI. *J Am Soc Nephrol, 24*(1), 37-42. doi: 10.1681/ASN.2012080800

72. Bhatraju, P. K., Zelnick, L. R., Herting, J., Katz, R., Mikacenic, C., Kosamo, S., Wurfel, M. M. (2019). Identification of Acute Kidney Injury Subphenotypes with Differing Molecular Signatures and Responses to Vasopressin Therapy. *Am J Respir Crit Care Med, 199*(7), 863-872. doi: 10.1164/rccm.201807-1346OC

73. Poston, J. T., & Koyner, J. L. (2019). Sepsis associated acute kidney injury. *BMJ, 364*, k4891. doi: 10.1136/bmj.k4891

74. Labelle, C, Marinier, A, Lemieux, S. (2019). Enhancing the drug discovery process: Bayesian inference for the analysis and comparison of dose-response experiments. *Bioinformatics. 2019 Jul; 35(14:i464-i473.*

75. Guo, X., Liu, B., Chen, L., Chen, G., Pan, Y., & Zhang, J. (2016). Bayesian Inference for Functional Dynamics Exploring in fMRI Data. *Comput Math Methods*

Med, *2016*, 3279050. doi: 10.1155/2016/3279050

76. Heaney, A. K., Alexander, K. A., & Shaman, J. (2019). Ensemble forecast and parameter inference of childhood diarrhea in Chobe District, Botswana. *Epidemics*, 100372. doi: 10.1016/j.epidem.2019.100372

77. Madan, H., Berlot, R., Ray, N. J., Pernus, F., & Spiclin, Z. (2019). Practical priors for Bayesian inference of latent biomarkers. *IEEE J Biomed Health Inform*. doi: 10.1109/JBHI.2019.2945077

78. Opstaele, L., Bielen, R., Bourgeois, S., Moreno, C., Nevens, F., Robaeys, G., & Van Vlierberghe, H. (2019). Who to screen for hepatitis C? A cost-effectiveness study in Belgium of comprehensive hepatitis C screening in four target groups. *Acta Gastroenterol Belg*, *82*(3), 379-387.

79. Shiao, C. C., Wu, P. C., Huang, T. M., Lai, T. S., Yang, W. S., Wu, C. H., . . . (CAKs), N. T. U. H. S. G. o. A. R. F. N. a. t. T. C. f. A. K. I. a. R. D. (2015). Long-term remote organ consequences following acute kidney injury. *Crit Care*, *19*, 438. doi: 10.1186/s13054-015-1149-5

APPENDICES

APPENDEX I

Institutional Review Board (IRB) Letter

The University of Arkansas for Medical Sciences IRB office provided approval letter to conduct the study. The IRB determined that the study was not human subject study.

UAMS
UNIVERSITY OF ARKANSAS FOR MEDICAL SCIENCES

Institutional Review Board
4301 West Markham, #636
Little Rock, AR 72205-7199
501-686-5667
501-686-7265 (fax)
http://irb.uams.edu/

FWA00001119

05/24/2018

PI Name: Baghal, Ahmad
PI Department: COM Biomedical Informatics
Number: 228146
Project Title: An Agile Research Data Repository Of Acute Kidney Injury Using Property Graph Databases

NOT HUMAN SUBJECT RESEARCH DETERMINATION

The Institutional Review Board Director or Designee reviewed your material and determined that this project is NOT human subject research as defined in 45 CFR 46.102, and therefore it does not fall under the jurisdiction of the IRB review process.

Committee Notes/Comments:

- The intent of this proposal is to develop an acute kidney injury (AKI) data repository by extracting de-identified data out of the Arkansas Clinical Data Repository. Therefore, this does not meet the regulatory definition of human subjects research and does not require IRB oversight.

Please keep the IRB advised of any changes that may require the project to be re-classified as human subject research.

If you have any questions, please contact an IRB administrator at 501-686-5667.
Click here to access study.

Ashley Block

UAMS IRB Administrator

APPENDEX II DEFINITIONS

Acute Kdiney Disease (AKI): a sudden episode of kidney failure or kidney damage that happens within a few hours or a few days. AKI causes a build-up of waste products in your blood and makes it hard for your kidneys to keep the right balance of fluid in your body. AKI can also affect other organs such as the brain, heart, and lungs

Enterprise Data Warehouse (EDW): is a database, or collection of databases, that centralizes a business's information from multiple sources and applications, and makes it available for analytics and use across the organization.

Graph database: Graph databases are NoSQL databases which use the graph data model comprised of vertices, which is an entity such as a person, place, object or relevant piece of data and edges, which represent the relationship between two nodes.

NoSQL Database: NoSQL databases are purpose built for specific data models and have flexible schemas for building modern applications. NoSQL databases are widely recognized for their ease of development, functionality, and performance at scale.

Database Schema: the skeleton structure that represents the logical view of the entire database. It defines how the data is organized and how the relations among them are associated. It formulates all the constraints that are to be applied on the data.

Bayes' Theorem: is a formula that describes how to update the probabilities of hypotheses when given evidence. It follows simply from the axioms of conditional

probability, but can be used to powerfully reason about a wide range of problems involving belief updates.

Bayesian Network: is a probabilistic graphical model that uses Bayesian inference for probability estimations, aim to model conditional dependence and causation by representing conditional dependence by edges in a directed graph.

Markov Chain: is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules.

Relational database: is a set of formally described tables from which data can be accessed in many different ways without having to reorganize the database tables.