Retrospective prediction – ER patients diagnosis & prognosis

**Lead Authors:** Dr. Kira Radinsky, Prof. Moshe Shoham, Jonathan Amir

**Data Extractor:**

**Statistics:**

**Epidemiologist:**

**Guarantor:**

|  |  |
| --- | --- |
| Protocol Template | |
| Retrospective prediction – ER patients diagnosis & prognosis | Title of research |
| Academic | Project type (Academic/Internal/Grant) |
| Technion  Dr. Kira Radinsky, Prof. Ran Balicer | Name of collaborating institution and contact person |
|  | Data transfer procedure (from external sources) |
| No Funding | Funding: Source and sum |
| [Retrospective Study] | Time period that funding covers |
| Research collaboration | If no funding, collaboration mechanism |
|  | Helsinki version |
|  | Helsinki (protocol version; number; and approval date) |
| no | Does the research have a prospective component (collecting new data)?  If yes, see **Appendix A** |

|  |  |  |
| --- | --- | --- |
| **Principal Investigator** | | |
| Name | Mobile | Email address |
| Dr. Kira Radinsky | 0544631808 | kiraradinsky@gmail.com |
| Address: | | |
| Co-investigator/s | | |
| Name | Mobile | Email address |
| Jonathan Amir | 0524500815 | Yonatanamir1@gmail.com |
| **Project Initiator** |  |  |
| Name | Mobile | Email address |
| Dr. Kira Radinsky | 0544631808 | kiraradinsky@gmail.com |

**Contents**

[ABSTRACT 4](#_30j0zll)

[BACKGROUND AND STUDY RATIONALE 5](#_1fob9te)

[Public health significance of research question 5](#_3znysh7)

[How this research fills gaps/ adds new evidence to the literature 6](#_2et92p0)

[Research question(s) 6](#_tyjcwt)

[Study objectives 6](#_3dy6vkm)

[METHODS 7](#_1t3h5sf)

[Context and Data 7](#_4d34og8)

[Study Design and data time period 7](#_2s8eyo1)

[Study population 7](#_17dp8vu)

[*Inclusion criteria- including continuous membership criteria* 7](#_3rdcrjn)

[*Exclusion criteria* 8](#_26in1rg)

[*Estimated sample size* 8](#_lnxbz9)

[Variables (Definitions, measurement of, and time periods) 8](#_35nkun2)

[*Outcomes* 8](#_1ksv4uv)

[1. Patient Discharge diagnosis 8](#_44sinio)

[2. Subsequent 30-day readmission post discharge diagnosis 8](#_2jxsxqh)

[*Main predictor(s) (socio-demographic, clinical, etc)* 9](#_z337ya)

[Statistical analysis (by study phase where relevant) 9](#_3j2qqm3)

[*a.*](#_1y810tw) *Descriptive statistics (tests for significance)* 10

[*b.*](#_4i7ojhp) *Multivariable modeling* 10

[Privacy 10](#_2xcytpi)

# ABSTRACT

**Rationale for conducting the study**

הקשר בין תלונות המטופל, הרקע ההיסטורי והסימנים החיוניים הראשוניים בכניסה למיון לאחר הטריאג לא נחקרו מספיק. במחקר זה נחקור איך ניתן לצפות את הדיאגנוזה בעזרת הנ״ל באופן מדויק באמצעות אלגוריתמים אוטומטיים. אם אכן תהליך זה יכול להיעשות באופן אוטומטי, זה ישפר בצורה דרמטית את זמן ההמתנה לחולים במחלקת המיון ועלול להפחית את השגיאות הרפואיות במיון.

**Study design and research phases**

אנו נבצע מחקר רטרוספקטיבי היסטורי של חולים שהוכנסו למיון בין התאריכים 1.1.2014-31.12.2015 עם 12 חודשי מעקב על מנת לעקוב אחר חזרות למיון. בנוסף לכל חולה נחקור את ההיסטוריה הרפואית של כל חולה עד 5 שנים לפני הביקור במיון.

המטרה היא להשוות את התוצאות של התחזיות האוטומטיות לאלו של אבחונים שנעשו על ידי רופאים.

יישום מודל החיזוי יתבסס על טכניקות למידה ממוחשבת אשר ילומדו באמצעות ההיסטוריה הרפואיית, האבחנה ופרוגנוזה. מודלים שונים ייבנו באמצעות טכניקות ML שונות: עצי החלטה, Boosted החלטות עצים וכו 'ו / או שינויים בנתונים שונים על מנת למצוא מודל צפוי על בסיס בדיקות צולבות היסטוריו.

* **Number of participants**

**~850,000 Clalit members visited an ER over 2014-2015**

* **Age range of participants**

גיל מעל 25 ופחות מ75 (כולל)

* **Sex of participants**

גברים ונשים

* **Inclusion and exclusion criteria**

במחקר ישתתפו החולים הבאים:

* גיל מעל 25 ופחות מ75 (כולל)
* ביקרו ב - ER באחד מבתי החולים בכללית בתאריכים 1/1/2014-31/12/2015 עם תלונה עיקרית על כאבים בחזה
* חברות בקופת חולים כללית של 5 שנים או יותר לפני קבלת ההרשמה

לא יכללו במחקר החולים הבאים:

* נשים בהיריון בתאריך האינדקס
* סרטן פעיל
* **Criteria for taking out participants from research**

מטופלים להם חסר אחד המרכיבים של הנתונים הרפואיים שהוזכרו לעיל

* **Dealing with special populations (not relevant for retrospective)**

לא רלוונטי (ניסוי רטרופספקטיבי)

* **Duration of research**

12 חודשים

# BACKGROUND AND STUDY RATIONALE

## Public health significance of research question

The rationale for conducting the study is based on serval elements related to the problematic ER patient intake methodology and systematic worsening of patient diagnosis process. Emergency rooms worldwide suffer from a few fundamental problems as a derivative of the growing and aging population: ED overcrowding, the rising costs of healthcare, the shortage of medical personnel in both developed and developing countries, and the unnecessary morbidity and mortality caused by human error. The acute problem of misdiagnosing patients (diagnostic errors - accounts for 37-55% of all ED errors, this includes misjudging symptoms, delayed diagnosis, incorrect diagnosis, and wrongful diagnosis) , based on a study conducted by "Johns Hopkins" In May of 2016 , is today the 3rd leading cause of death in the US (after cancer and chronic respiratory disease).(1,3,4)

The goal of the research and the system that is being developed, which is designated for hospital use, is to constitute a new type of auxiliary medical resource that can potentially provide faster, more comprehensive and more accurate diagnosis for patients in the ED. Using an automated system, a set of sensors and AI/ML algorithms will compare a given patient’s status in the ED with large and constantly evolving database of medical indicators, diagnosis, treatment and outcomes data, the system provides a tool that would massively reduce the cost, time, and error rate of the initial steps in the diagnosis-to-treatment cascade.

As chest pain is the second most reported (2) complaint in the ED, we wish to focus our study on the prediction of the diagnosis of the patient with chest pain in the entrance at the ED and compare the model's performance to that of the decision making process of medical doctors. We hypothesize that the use of advanced machine learning techniques will yield adequate results that can be used to automate and support the ED processes.

The significance of the pure retrospective research is to constitute a proof of concept to the predictive model and the algorithmic/AI components of the system. The model will be based on pure historical data (real time data will not be included in this stage) in order to create a prediction regarding the final diagnosis and prognosis of the ER patient. The prediction will be based on the following historical medical data: the patient's demographic information, the patient's clinical EMR, data from the "Ofek" system (the Israeli national database), patient's chief compliant (as registered in the EMR), the patient's historical ER visits, the patient's historical admissions, the patient's drugs/medicines (as registered in the EMR), and ambulatory diagnostic and treatment visits. The system is initially intended for hospital's Emergency Rooms ("**ER**/**ED**") but can be expanded as well to community care centers and retail clinic centers.

## How this research fills gaps/ adds new evidence to the literature

We wish to apply the machine learning techniques to build a DSS (Decision Support System) for facilitating quick and accurate diagnoses in the ED/ER and validate the findings (predicted diagnosis and prognosis of the patient and insights regarding his diagnosis across varying patient profiles) based on the historical medical data. No research such as this has been conducted in Israel to date.

The DSS market has been constantly growing in the past few years due to the growing use of big data analytics and A.I. systems that help doctors diagnose patients. A DSS requires the doctor to perform the necessary tests and enter the indexes in order to generate a diagnosis and this process is one of the reasons 70% of ED doctors' time during a standard shift is spent on working in front of a computer rather than with patients. The research and the system that is being developed offers an end-to-end solution which creates a fully automated patient ED intake pipeline that automates monotonous and repetitive actions performed by the ED doctors. To the best of our knowledge, such DSS systems were not performed in hospitals.

## Research question(s)

1. What are the parameters (patient demographics, clinical data and historic medical data as documented in the EMR) and parameter combinations associated with the likelihood of different diagnoses based on the patient chief complaint of chest pain.

2. Can diagnoses and prognoses be predicted automatically without human involvement?

## Study objectives

1. To identify the Clalit patient population that are admitted to or discharged from Clalit hospital ERs with chest pain and assess the combinations of different diagnoses based on this patient-specific chief complaint.

2. To determine if diagnoses and prognoses be predicted automatically without human involvement by building a predictive diagnostic model for Emergency Rooms that could be viable with a high accuracy rate.

3. To apply the machine learning techniques to build a DSS (Decision Support System) for facilitating quick and accurate diagnoses in the ED/ER

# METHODS

## Context and Data

Clalit Health Services (CHS), the largest integrated health care service provider and payer system in Israel, has over 4 million active members. CHS has a comprehensive health care data warehouse which combines hospital and community medical records, laboratory and imaging information, pharmaceutical records, health care costs, and Ministry of the Interior vital statistics of all the members. Membership turnover within Clalit is less than 1% annually, facilitating the study of population trends over time.

## Study Design and data time period

This is a population-based retrospective cohort study. The cohort will include all patients who visited the ER at any CHS hospital between January 1, 2014 and December 31, 2015 (the study period). The date of an individual's first admission during the study period will be their cohort entry and defined as the index date. We will follow 30-day readmission patterns of all individuals following their index hospitalization for up to one year.

We will have an additional six months of follow up to evaluate hospital readmission among those who were discharged in December of 2016.

Background historical data will be collected for 5 previous to each member's individual index date.

## Study population

### *Inclusion criteria- including continuous membership criteria*

All individuals with continuous membership in Clalit for 5 years previous to index date, and aged 24 up to age 75 at index date.

At this stage, the study will only include patients whose first ambulatory visit in the ER in Clalit hospitals during the study period was due to the complaint of "chest pain" as documented on admission or discharge.

### *Exclusion criteria*

* Pregnant women in the index date
* Active Cancer

## *Estimated sample size*

This population will draw from the over the 850,000 members of CHS who, on average, are admitted into the ER over a two-year period.

## Variables (Definitions, measurement of, and time periods)

The prediction will be based on the following historical medical data: the patient's demographic information, the patient's clinical EMR, data from the "Ofek" system (the Israeli national database), patient's chief compliant (as registered in the EMR), the patient's historical ER visits, the patient's historical admissions, the patient's drugs/medicines and ambulatory diagnostic and treatment visits.

Data will be taken from the inpatient and outpatient clalit data warehouses. Socio-demographic information will be taken as of the index date or from current demographics where historical data is not available. Medical history of the following variables will be taken from outpatient visits, inpatient admissions, chronic registry, oncology registry, medications, imaging, clinical markers, and laboratory values when necessary, as of the index date and the end of follow-up.

### *Outcomes*

# Patient Discharge diagnosis

The patient diagnosis at admission and discharge as coded in the ER hospitalization, based on predetermined classes of diagnoses (that will be created during the ML training phase) using [ICD9](https://www.google.co.il/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjFuuW7rpLVAhUIHxoKHS48De4QFggmMAA&url=http%3A%2F%2Ficd9.chrisendres.com%2F&usg=AFQjCNFtkNvuzOJXQfNdOaAOG_7RUqLs6Q) or [SNOMED](http://www.snomed.org/snomed-ct).

# Subsequent 30-day readmission post discharge diagnosis

This is defined as all hospitalizations within 30-days post discharge of index hospitalization within the year following the index hospitalization. The rationale is to identify misdiagnoses and to train the predictive model on the last final diagnosis.

### *Main predictor(s) (socio-demographic, clinical, etc)*

The following variables in the past 5 years (including all repeat records, where applicable):

1. Socio-demographics
   1. Sex
   2. Age
   3. Socioeconomic status by clinic and by address
   4. Supplementary insurance status
   5. Country of birth and immigration date
   6. Ethnicity by country of individual’s or parents’ birth
   7. Sector (clinic level data - predominantly Arab / Jewish)
   8. Clalit affiliation by district, sub-district and clinic
2. Clinical markers/Co-morbidities
   1. Smoking status
   2. BMI, height, weight
   3. Charlson co-morbidity index
   4. Johns Hopkins Adjusted Clinical Groups (ACG) score
   5. Chronic diseases (all documented in 5-years pre index date), for example:
      * History of malignancies and active malignancy
      * Cardiovascular diseases (ischaemic heart disease, cerebrovascular disease).
      * Framingham risk score/ SCORE
      * Diabetes (taken from CRI registry)
      * Hypertension (taken from CRI registry)
      * Asthma
3. All chronic medications
4. All laboratory tests
5. All imaging tests/results
6. All vitals data at index and readmission hospitalizations
7. All anamnesis data of the triage

## Statistical analysis (by study phase where relevant)

# *Descriptive statistics (tests for significance)*

The distribution of the baseline demographic and clinical characteristics as well as the outcomes across the study group will be examined. We will also present the univariate associations between the socio-demographic and clinical covariates and the possible diagnostics outcomes, using ANOVA test for normal continuous variables, Kruskal Wallis test for ordinal/not normally distributed continuous variables and Chi-square for categorical ones.

# *Multivariable modeling*

Second, we intend to build several machine learning models, including boosted trees, SVMs and Neural Networks, for each possible diagnosis of chest pain. The output variable will be the diagnosis at the latest discharge, if multiple hospital re-admissions were made in the years after the index discharge. Each person will be considered only once.

***Statistical program used such as:*** Data will be extracted using MS-SQL. Analysis will be performed using python and scala

# Privacy

Data extraction and analyses will be conducted at the Clalit Research Institute (CRI) by employees of the CRI. The raw data extracted are coded, viewed and stored only within the CRI. Once data are analyzed and leave the CRI they do not contain any identifiable information.