VERSION 1.0

DATA MANAGEMENT PLAN

Updated by:
Amin Abuedieh
Abed
Fatima Twam
Saleh Abbas
Kayan Abukhaizaran

Data Management Plan

IONREMEDIES

TABLE OF CONTENTS

1. DATA DESCRIPTION, COLLECTION AND REUSING EXISTING DATA

A. DATA DESCRIPTION

The data were obtained from the Health Information System (HIS) of Istishari Arab Hospital (IAH). The following variables were requested:

Diagnosis

Admission and Discharge Date and time

Gender

Age

Height

Weight

Medication Orders and Medication History

Observations: Blood Pressure, Heart Rate, Respiratory Rate , Temperature

Past Medical History:

Past Surgical History

Smoking and Alcohol consumption

Lab Tests : CBC, Liver Function Tests, Lipid Profile

1. **Reusing Existing Data**

   The data obtained were for patients' visits for the year 2021, we excluded the outpatient and emergency department visits, as our study will focus primarily on inpatients with > 24 hours hospital stay. We were provided with four excel sheets with the following datasets; Diagnosis, Vitals, Labs and Medication datasets, all composing the requested variables. The diagnosis dataset consisted of 15882, the

vitals contained 122377, the labs had 1011250 records and the medications had 510270 records.

The variables (smoking history, past medical history, past surgical history, and medication history) were not available as requested, thus we were provided with the first physician note for all the patients. We needed to extract the needed information from the physician notes.

To link the four different datasets, the patient encounter ID was the common variable between the datasets.

2. **Data Collection**

To ensure the continuity of collecting the data, the IT team at IAH created a query so that the data will be provided to our team on a monthly basis.

B. PURPOSE

The purpose of collecting the following variables is stated below:

Diagnosis

- To study patients admitted to the hospital with hypertension related complications.

-To capture the effect of the current diagnosis with the medication plan.

Admission date and time

-To calculate the length of stay for each patient.

Gender

-To differentiate between male and female based on normal ranges.

Age

-To classify our results based on age groups.

Height and Weight

-To study the correlation between higher BMI, hypertension disease and the effect of the doses.

All the Medication Orders and Medication History

-To study in depth antihypertensive medications.

-To document the time of administration and reflect it on the observations.

-To study the impact of medication history on current medication plans.

Observations and their read time: Blood Pressure, Heart Rate, Respiratory Rate , Temperature

-To study the direct effect of medication on the observations.

Past Medical History:

-To take into account the interference of previous medical history on the metabolism of the selected medications, in addition to the baseline for the body observations.

Past Surgical History

-To take into account the previous surgeries performed.

Smoking and Alcohol consumption

-Smoking and alcohol consumption alters the body's observations as well as drug metabolism.

Lab Tests : CBC, Liver Function Tests, Lipid Profile

-To take into account the mentioned lab tests on the medication metabolism.

C. DATA TYPE

The data provided were of the following types.

Diagnosis Dataset

```
[ ] DAG.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 15882 entries, 0 to 15881
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   #                        15882 non-null  int64
 1   Encounter ID             15882 non-null  float64
 2   Patient ID               15882 non-null  float64
 3   Gender                   15882 non-null  object
 4   Age                      15882 non-null  float64
 5   Encounter Type           15882 non-null  object
 6   Current_Department_Name  15882 non-null  object
 7   Diagnosis                15880 non-null  object
 8   First Physician Note     15788 non-null  object
 9   Amission Date            15882 non-null  datetime64[ns]
 10  Discharge Date           15875 non-null  datetime64[ns]
dtypes: datetime64[ns](2), float64(3), int64(1), object(5)
memory usage: 1.5+ MB
```

Diagnosis Dataset Data type

In all datasets, the Encounter ID and Patient ID values were converted from float to integer. Regarding the Admission and Discharge Date, the data was of datetime type, and because the time is an important variable in our study, we segregated the datetime into two columns one containing the time and the other the date. We also calculated the length of stay for each patient using the date column.

The results of our study cannot be generalized to the pediatric population, thus we excluded all encounters for patients who are under 18 years old.

To ease the handling of text values, the Diagnosis variable was mapped with the ICD-10 coding scheme.

Using Regular expression, we were able to extract the needed variables from the first physician note.

Example:

```
[55] pmh=0
     for x in DAG['First Physician Note']:
       gg=str(x)
       if ('past medical history' in gg.lower()) or ('past medical' in gg.lower()) or ('medical history' in gg.lower()) or ('pmh' in gg.lower()) or ('pmhx' in gg.lower()):
         pmh=pmh+1
     print(pmh)

     7284
```

For future purposes, we communicated with the IT department at the IAH hospital. Our requested variables will be obtained from new screens currently in the testing phase, as shown in Annex 1.

Vitals Dataset

```
[ ] Vital.info()

    <class 'pandas.core.frame.DataFrame'>
    Int64Index: 122377 entries, 0 to 122376
    Data columns (total 11 columns):
     #   Column            Non-Null Count   Dtype
    ---  ------            --------------   -----
     0   ENCOUNTER_ID      122377 non-null  float64
     1   PATIENT_ID        122377 non-null  float64
     2   READ_DATE         122377 non-null  datetime64[ns]
     3   HEIGHT            122377 non-null  float64
     4   WEIGHT            122377 non-null  float64
     5   TEMP              78218 non-null   float64
     6   PULSE             86567 non-null   float64
     7   RESPIRATORY_RATE  122377 non-null  float64
     8   RES_RATE          11018 non-null   float64
     9   BP_SYSTOLIC       122377 non-null  float64
     10  BP_DIASTOLIC      122377 non-null  float64
    dtypes: datetime64[ns](1), float64(10)
    memory usage: 11.2 MB
```

Vitals Dataset Data type

During the data cleaning phase, we encountered many outliers and fault values due to human errors. We provided the IAH with validation tools to be applied on the observation fields; among our suggested validation tools is that; the Pulse,BP, Pulse O2_Sat to be of an integer type.

Labs Dataset:

```
[ ]  Lab.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1011250 entries, 0 to 1011249
Data columns (total 13 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   #                 1011250 non-null  int64
 1   Encounter ID      1011250 non-null  int64
 2   Order Type        1011250 non-null  object
 3   Patient ID        1011250 non-null  int64
 4   Order ID          1011250 non-null  int64
 5   Result Category   1011250 non-null  object
 6   Product Name      1011250 non-null  object
 7   Result Name       1011250 non-null  object
 8   Result Notes      1011250 non-null  object
 9   Normal Range      893704 non-null   object
 10  Unit              960930 non-null   object
 11  RESULT_DATETIME   1011250 non-null  datetime64[ns]
 12  APPROVE_DATETIME  1011250 non-null  datetime64[ns]
dtypes: datetime64[ns](2), int64(4), object(7)
memory usage: 100.3+ MB
```

Labs Dataset Data type

The Order ID column and the Order Type columns were dropped from the dataset as they are irrelevant to our study.

Medications Dataset:

```
[ ]  Med.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 510270 entries, 0 to 510269
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   EncounterID        510270 non-null  int64
 1   Patient ID         510270 non-null  int64
 2   Product Name       510270 non-null  object
 3   Instructions       502883 non-null  object
 4   Order Date Time    510270 non-null  datetime64[ns]
 5   Applied Date Time  510270 non-null  datetime64[ns]
dtypes: datetime64[ns](2), int64(2), object(2)
memory usage: 23.4+ MB
```

Medication Dataset Data type

To ease the handling of the medications, the product name column was mapped with the product ID which was provided to us by the IAH. Hence, the product name is now of an integer type.

2. DOCUMENTATION AND DATA QUALITY

To ensure the quality of the existing data, we ran various cleaning and validation methods on the selected variables.

We replaced the outliers and null values with 0.

Example:

- Detecting the outlier for the Temperature variable.

```
[55] pmh=0
     for x in DAG['First Physician Note']:
       gg=str(x)
       if ('past medical history' in gg.lower()) or ('past medical' in gg.lower()) or ('medical history' in gg.lower()) or ('pmh' in gg.lower()) or ('pmhx' in gg.lower()):
         pmh=pmh+1
     print(pmh)

     7284
```

Detecting outliers for Temperature variable

- Replacing the outliers with 0.

```
[ ]  Lab.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 1011250 entries, 0 to 1011249
     Data columns (total 13 columns):
      #   Column            Non-Null Count    Dtype
     ---  ------            --------------    -----
      0   #                 1011250 non-null  int64
      1   Encounter ID      1011250 non-null  int64
      2   Order Type        1011250 non-null  object
      3   Patient ID        1011250 non-null  int64
      4   Order ID          1011250 non-null  int64
      5   Result Category   1011250 non-null  object
      6   Product Name      1011250 non-null  object
      7   Result Name       1011250 non-null  object
      8   Result Notes      1011250 non-null  object
      9   Normal Range      893704 non-null   object
      10  Unit              960930 non-null   object
      11  RESULT_DATETIME   1011250 non-null  datetime64[ns]
      12  APPROVE_DATETIME  1011250 non-null  datetime64[ns]
     dtypes: datetime64[ns](2), int64(4), object(7)
     memory usage: 100.3+ MB
```

For the ongoing data collection phase, our team reviewed the literature for ranges of the variables, and suggested validation tools to be applied by the IT department team at IAH.The tools contained constraining minimum and maximum values and allowing one type of data entry for certain fields.

The proposed validation tools will be applied, thus ensuring better data quality for the future.

3. STORAGE AND BACKUP DURING THE RESEARCH PROCESS

- All our data will be stored on Cloud service; we will be considering Amazon Web Service (AWS) or Google Cloud Service. Through Site to Site VPN, the integration between our and the IAH server will be managed.
- Daily backup for the data will be carried, and will be stored on a different cloud storage server.
- The transferred data between the hospital and our cloud server will be encrypted through VPN.
- Our server will be protected by a firewall.
- We will use integrity checks to validate that the backup data is valid.
- We will consider the HL7 standard to manage the security of health data.
- Our server will be protected by an Anti Malware; Endpoint Detection and Response System (EDR).
- Only authorized users will have access to our system.

4. LEGAL AND ETHICAL REQUIREMENTS, CODE OF CONDUCT

We guaranteed confidentiality by assuring that the information is not available for anyone who was not involved in the study. The patient and healthcare provider names were not required or documented. The patient ID will be stored in a hash value to avoid reverse engineering the identity of patients.

The approval from the IAH ethics committee and administration was obtained, as shown in Annex 2.

5. DATA SHARING AND LONG-TERM PRESERVATION

We will use cloud services and data will be protected against unauthorized users.

There is no need to destroy the data since there is no contractual liability.

Data provided to us by the hospital will be used to build our model and for research purposes only. Some of the results will be communicated back to the hospital.

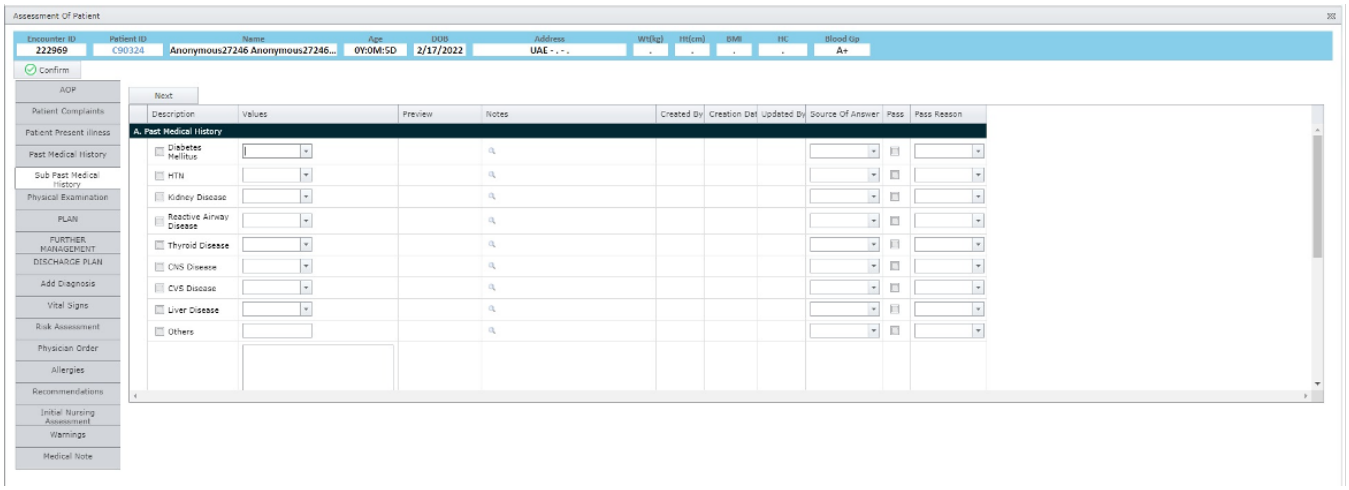
Up to the moment we are using the Google Collab, which is a known and reliable tool.

6. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES

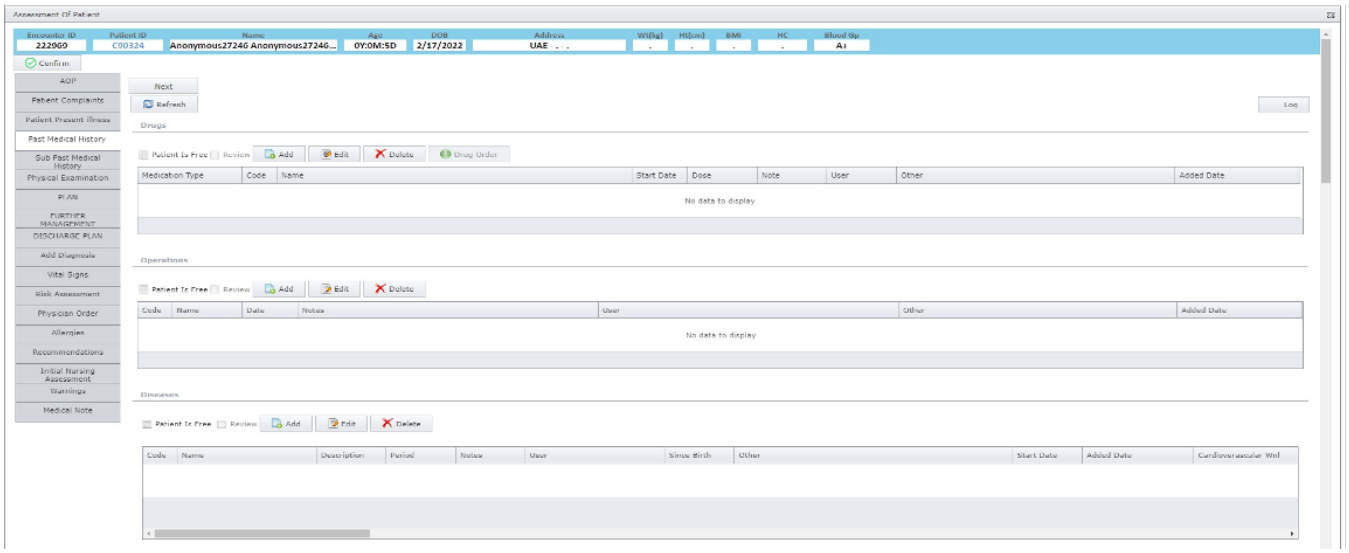
- The domain experts team (Kayan Abukhaizaran, Fatima Twam and Abed) will handle the data collection process, metadata production and ensure the data quality.

-The IT experts (Saleh Abbas and Amin Abueideh) will handle building the model, analysis, storage and backup, data archiving, and data sharing.

ANNEX 1



Past Medical History Screen



Past Medication and Surgical History Screen

Smoking History Screen

ANNEX 2

المستشفى الإستشاري العربي
ISTISHARI ARAB HOSPITAL
أكثر من مجرد مستشفى

IAH Research Application Form

Date	19/01/2022
Name of investigator	Kayan Abukhaizaran
Mobile No.	0599133218
Email	kayan@iah.ps
Expected start date	
Expected completion date	
Name of Company/University	Birzeit University
Attached needed	
Investigator CV	☐Yes ☐No
Study Proposal	☐Yes ☐No
Consent Form	☐Yes ☐No
Data Collection Tools	☐Yes ☐No
Informed Consent (Arabic & English)	☐Yes ☐No
For COO Office	
Receiving Date	
Application completed	☐Yes ☐No
COO Director Note	
Transfer Date	
COO director Sig.	
For Ethical Committee	
Receiving Date	24/1/2022
Ethical Committee Approval	☑Yes ☐No
Ethical Committee Note	Analytical research for non medical purposes from ethical po... approved
Head of Ethical committee Sig.	25/1/2022
CEO Note	OK
CEO Sig.	26/1/2022

- For Non Experimental Research only