Aurora Anticoagulation and Anticlotting's EMRs: Extraction, Transformation and Loading

Kourosh Ravvaz

July 7, 2013

Contents

1 Introduction

The goal of this effort is to extract warfarin treated patients from the Aurora Health Care (AHC) patient database which will be used to create a statistical characterization and use a portion of the patient data to train a Bayesian Network Model to produce accurate Aurora patient population clinical avatar populations. The clinical avatars are used to produce pharmacogenomic simulations to test the predicted accuracy and efficacy outcomes against Aurora Health Care outcome data.

2 The Process of Extraction, Transformation, Loading

The Aurora Health Care (AHC) project team mined all project data from the Aurora Health Care hospitals and clinics data warehouse for the period of 2002 to 2011. The patient data was de-identified by the AHC project team before distribution to the UWM team. The UWM team extracted, transformed, and then loaded the de-identified subject records into a working database. The cohort includes patients with evidence of prescription of any of the following medications: Coumadin (Warfarin), Heparin, Ticlopidine (Ticlid), Clopidogrel (Plavix), Dipyridamole (Persantine), Abciximab (ReoPro), Eptifibatide (Integrilin), Tirofiban (Aggrastat), or Dabigatran (Pradaxa). This effort resulted in a total of 157,450 de-identified data records. Each data record includes (as available): gender, race, height, weight, age, day of visit, patientś zipcode, patientś city, providerś zipcode, smoking status, INR result, medications received (day, dose, frequency), interacting medications (Amiodarone, Simvastatin, Fluvastatin, Lovastatin, Atrovastatin, Rosuvastatin, Pravastatin, Aspirin), medication indications (by ICD-9 codes: Orthopedic surgery (hip or knee), Deep vein thrombosis, Pulmonary embolism, Atrial fibrillation, Atrial flutter, Atrial fibrillation and flutter, Stroke, Heart valve replacement) and medication adverse events (by ICD-9 codes: Deep vein thrombosis, Pulmonary embolism, Stroke, Myocardial infarction, Bleeding).

The original Aurora Anticoagulation Anticlotting Raw dataset (ARD) was delivered in the MS Access format consisting of the following 12 tables:

1. PATIENT AGE
2. PATIENT GENDER
3. PATIENT HEIGHTS
4. PATIENT RACES
5. PATIENT RESIDENCES
6. MEDICAL INDICATIONS
7. INTERACTING MEDICATIONS
8. ADVERSE EVENTS
9. MEDICATIONS
10. INR
11. PATIENT PROVIDER INFO
12. PATIENT WEIGHTS
13. PATIENT TOBACCO USAGE

After a rigorous iterative process of cleaning, transformation and data quality control and assurance, the data was loaded into a working database now called **WiAD** standing for "Wisconsin Anticoagulation Anticlotting Database". An interactive data profiling and population "segmentation" tool **WiAD-Miner** was developed in R using RStudio and used to facilitate the process of identifying those segments of the total population that satisfied inclusion criteria such as medication name, start-end date of treatment, geographical location (city and county), and number of visits. WiAD-Miner provides a user friendly interactive interface to subset the entire WiAD population and data records and presents a clear view of each extracted subset by producing statistical characteristics and visual profiling (e.g. trend of dose changes during the desired period). The extracted subsets are available for download to conduct further analyses using other tools such as SAS.

Through the following sections, the process of data cleaning, transformation, and quality control are explained.

2.1 Data Transformation and Quality Control

The process of data cleaning, data quality control and quality assurance in the context of Aurora electronic medical records (EMR) data reuse for research involves several iterative phases of data analysis and transformation. Two related approaches for data analysis were taken; a.) data profiling and b.) data mining. These approaches provide a statistically sound assessment of different aspects of quality at attribute and cross-attribute levels such as completeness, correctness and concordance.

a.) Our data profiling analysis is conducted in three steps. **First**, the method focuses on individual attributes (fields) in the provided dataset. This step provides a statistically sound assessment of aspects of quality of each attribute including identification of various characteristics of the data such as "type", value range, discrete values and their frequency, distribution and variance, uniqueness, occurrence of null values, and the pattern of free text values (e.g., values under "medication name"). **Second**, the results derived in the first step are used to handle issues on completeness (e.g., missing values at random, missing values not at random, and partial and complete duplications) by methods such as taking a "complete case strategy" limiting the analysis to patients with complete information for all variables, a simple deletion of all incomplete observations and or imputation and to handle issues on correctness (e.g., misspellings, awkward or inaccurate abbreviations, contradictory values, values outside of range, and unexpected changes of sequential data over time) by methods such as transforming and normalizing data elements and prepare them for integration using a data dictionary and looking for elements with values that are outside biologically valid or plausible ranges or that changed questionably or implausibly over time or zero valued elements. **Third**, concordance assessments are done by; a.) Agreement analyses between elements within attributes and b.) Distribution analyses of attributes within the EMR with that of the same information from reference distributions such as similar medical practices, studies or national or state rates (e.g. AHC patient geographical distribution against statewide distributions of race and gender) using both statistical tests and GIS visualizing tools (such as Google Map). Table 1 depicts examples of data problems captured by our data profiling process.

b.) Our robust data mining process also identifies specific data patterns in the data set. Given that this is a temporal and longitudinal dataset, it is very important and useful to take advantage of descriptive data mining models to explore the data more deeply. The models such as clustering, summarization, association discovery and sequence discovery are of interest and applied in this case to explore and evaluate, for example, the pattern of treatment, level of anticoagulation control, and outcome of treatment (e.g., stroke, thromboembolic events, bleeding events, hospitalization, patterns of INR monitoring, and time in the therapeutic range) with different anticoagulation anticlotting agents in patients with different indications.

Scope/Problem		Original Data
Attribute	Missing values	Race = "UNKNOWN"
	Misspellings	Medication Name= "warfin"
	Awkward Abbreviations	Medication Frequency= "Q3WK"
	Free Text Embedded values	Medication Name= "warfarin 2.5 mg 5 days a week and 2 mg two days a week"
	Miscoded values	Patient zip code= "WI"
	Incorrect values	Weight= -165
Record	Violated attribute dependency	City= "Milwaukee", zip code=99999
	Duplicated records	ID=165, Day= 199; Medication= Warfarin 4 mg, Frequency= QOD; ID=165, Day= 199; Medication= Warfarin 2 mg, Frequency= daily
	Contradicting records	ID= 213, Adverse Event ICD 9 = 434.91, Day= 1; ID= 213, Adverse Event ICD 9 = 435.9, Day= 1 / ID= 78, Day= 1101, Medication= Coumadin, Dose= 3 mg; ID= 78, Day= 1101, Medication= Coumadin, Dose= 4 mg

Figure 1: Examples for data cleaning and quality control problems.

3 General Cleaning

In this section, the first round of data cleaning is conducted by taking the following steps on the demographic data:

- First, to identify and exclude records whose attributes' values are missing,
- Second, to identify and exclude complete duplicated records,
- Third, since in the final modeling and simulation study each subject should have one single value for each demographic characteristics (e.g., weight, height) and given that each subject should have been almost stable in most of the demographic characteristics during the treatment period at AHC, we have to select an aggregation function for obtaining a representative measure of longitudinal records of demographics,
- Fourth, to refine the demographic data by excluding highly likely outliers.

3.1 Age

1 subject(s) has missing age information. 2662 subjects have age of 0. Subjects with age of zero or missing age are excluded from the entire study. black list has 2663 and useable list has 154787 Here is the age distribution of the whole ARD population after excluding records with missing age and age of 0.

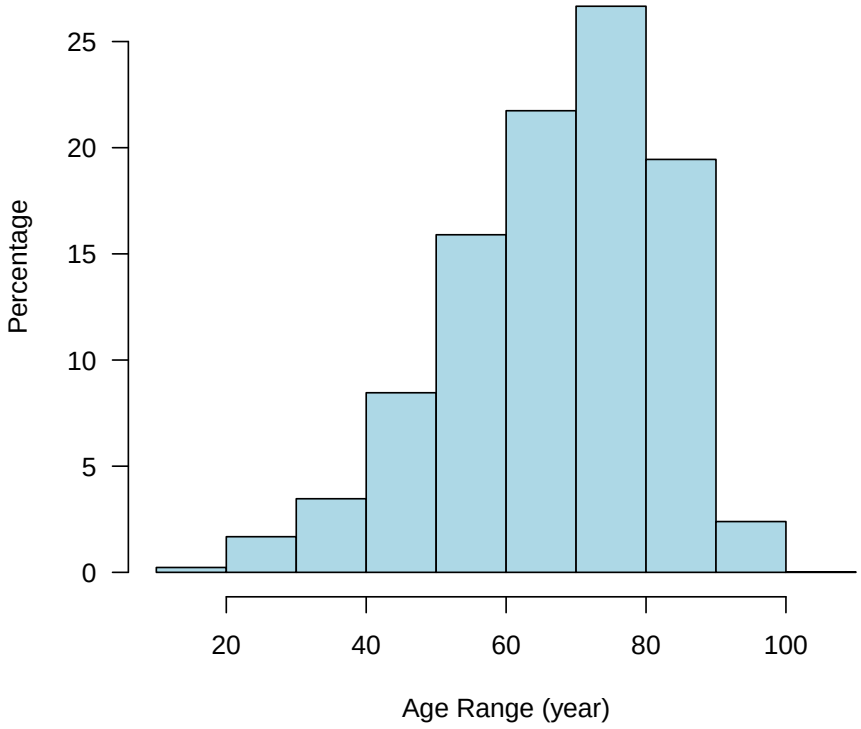

Figure 2: ARD Population - Age Distribution

Age.Range	Percentage
<18	0.23
18-24	0.86
25-34	2.46
35-44	5.81
45-54	14.11
55-59	9.92
60-64	10.62
65-74	22.22
75-84	22.70
>85	11.08

Table 1: Age Distriubtion of the Aurora Raw Data

The number of subjects in whole poulation with an age under 18 is 357 . **The subjects with age of 18 or lower are excluded from the study population as our study's age inclusion criteria is age on 18 and over.**

black list has 3020 and useable list has 154430

3.2 Gender

16 subjects have missing gender information. Here is the gender distribution of the ARD (Table 4) and also that of Milwaukee County and the State of Wisconsin (Table 5) for comparison.

Gender	Percentage
Male	50.34
Female	49.65
Missing	0.01

Table 2: Gender Distriubtion in the Aurora Raw Data

	MKE(%)	WI(%)
Male	48.30	49.60
Female	51.70	50.40

Table 3: Gender Distriubtion in Milwaukee County (MKE) and Wisconsin (WI).

The subjects with missing gender are excluded from the study population.

black list has 3036 and useable list has 154414

3.3 Race

In the ARD, each subject has one race record. Race has the following attributes: SURROGATE_ID, RACE, SOURCE_SYSTEM.

Since race is a required information for the future analysis, I have to check to make sure if there is any subject with missing race information and if so exclude them.

Race types are: White, Black or African American, UNKNOWN, Asian, American Indian or Alaskan Native, Native Hawaiian/Other Pacific Islander1662 subjects have missing race information. 52.85 percent of original population have race information. The following table depics the race distribution of the ARD population.

black list has 4559 and useable list has 80885

Race	Percentage
White	90.41
Black or African American	8.52
Asian	0.79
American Indian or Alaskan Native	0.27
Native Hawaiian/Other Pacific Islander	0.02

Table 4: Race Distriubtion in the Aurora Raw Data

3.4 Height

In the ARD, each subject has multiple height records. Height has the following attributes: SURROGATE_ID, EFFECTIVE_DAY, HEIGHT, SOURCE_SYSTEM

For the height records, a few steps are taken:

- First, to identify and exclude height records whose HEIGHT attributes are missing.
- Second, to identify and exclude duplicated height records.
- Third, since in the final modeling and simulation study each subject should have one height record and given that each subject should have been almost stable in height during the treatment period at AHC, we take the median of each subject's height records and take it as the representative height of the subject.
- Fourth, the height medians are refined by excluding the ones that here are considered outliers (i.e., median heights over or lower than 3 standard deviations).

Heights records include 333385 duplicate records.The ARD poulation on average has 3.07 height records/subject.The average height of the original poulation is 66.09 inch. 938 have outlier height. black list has 4612 subjects and useable list has 63550 subjects.

3.5 Weight

In the ARD, each subject has multiple weight records. Weight has the following attributes: SURROGATE_ID, EFFECTIVE_DAY, WEIGHT, SOURCE_SYSTEM

For the weight records, a few steps are taken:

- First, to identify and exclude weight records whose WEIGHT attributes are missing.
- Second, to identify and exclude duplicated weight records.
- Third, since in the final modeling and simulation study each subject should have one weight record and given that each subject should have been almost stable in weight during the treatment period at AHC, we take the median of each subject's weight records and take it as the representative weight of the subject.
- Fourth, the weight medians are refined by excluding the ones that here are considered outliers (i.e., median weights over or lower than 3 standard deviations).

Weights records include 61312 duplicate records.The ARD poulation on average has 7.39 weight records/subject.The average weight of the original poulation is 195.12 pound. 79 have outlier weight. black list has 4668 subjects and useable list has 62504 subjects.

4 General Overview of the Medications Records

The Auroa Raw Dataset (ARD) includes 2,349,633 longitudinal medication records representing 157,450 unique subjects. As a result of being a longitudinal dataset, each subject has more than one medication record.

4.1 Medications Records' Attributes

Each medication record has the following attributes:

- SURROGATE_ID
- DAY_PRESCRIBED
- MEDICATION_NAME
- FREQ, DOSE_QTY
- DOSE_QTY_UNIT
- SOURCE_SYSTEM

4.2 Distribution of Medications Records

First, I have to identify any specific medication cohort by extracting any records in the Aurora Raw Dataset (ARD)'s Medications records including any of the following terms:

- Warfarin or Coumadin
- Heparin
- Clopidogrel or Plavix
- Dabigatran or Pradaxa
- Eptifibatide or Integrilin
- Abciximab or ReoPro
- Ticlopidine or Ticlid
- Tirofiban or Aggrastat
- Dipyridamole or Persantine

After identifying the records, they have to be cleaned and transfomred for the following issues:

- Duplications
- Free Text Embedded values

The free text embedded values of the medication cohorts are used to extract dosage, unit, and route.

Then the next step is to take a strategy on how to deal with the incomplete records as we need records including name of the medication, day prescribed, frequency of the medication, dosage of the medication, unit, and probably medication route.

Here is the medications records distribution by medication.

99.02% of the medications records includes one of the three medications of Warfarin (37.85%), heparin (36.24%) and clopidogrel (24.93%). Warfarin records which have dosage information have the

	Records with med. name(%)	Subjects with med. name(#)	Records with med. name & dose(%)	Subjects with dose(#)
Warfarin	37.85	74102	11.00	51896
Heparin	36.24	71537	2.31	23899
Clopidogrel	24.93	61517	5.17	41851
Dabigatran	0.30	1793	0.13	1599
Dipyridamole	0.27	2192	0.05	525
Eptifibatide	0.19	3066	0.15	2886
Ticlopidine	0.14	434	0.04	310
Abciximab	0.08	1310	0.05	1181
Tirofiban	0.00	48	0.00	0
Total	100.00	215999	18.90	124147

Table 5: Medications Records Distriubtion by Medication.

highest percentage of whole the records (11%). Although the number of heparin records is bigger than the number of clopidogrel records, the percentage of clopidogrel records with dosage information out of whole the records is higher than that of heparin. The above table indicates that a big number of subjects have received more than one medication. The following table shows the number of subjects who have been under treatment with 1 to 6 medications.

Number of Medications	1	2	3	4	5	6
Number of Subjects	78613	17182	3185	376	21	1

Table 6: Number of Subjects by Number of Received Medications - Only medication records with dosage information included.

5 Warfarin Cohort

Given that each medication has specific characteritsics, the analysis of each medication is done separately after general anaysis done on all of the records. Warfarin cohort has the biggest number of records in the ARD and is also the medication of interest for the modeling and simulation study. So, the analysis of this cohort's records is done first.

5.1 Warfarin Cohort's Demographics

The warfarin cohort implies warfarin and coumadin records with at least one piece of dosage information.

```
The warfarin cohort includes 186 records which have more than one embedded digit
(e.g., warfarin 2.5 mg 5 days a week and 2 mg two days a week).
To take advanatage of these records, they should be manually cleaned and transformed.
```

The warfarin cohort's identification numbers are used to extract demographic information of the cohort. One important point to note is that, although the ARD is a longitudinal dataset, there is **only one Gender and one Age record** for each unique subject.

5.1.1 Age

Here is the age distribution of the warfarin cohort after excluding records with missing age and age of 0.

Age.Range	Percentage
<18	0.00
18-24	0.73
25-34	2.19
35-44	5.02
45-54	11.55
55-59	8.42
60-64	9.86
65-74	23.97
75-84	26.87
>85	11.40

Table 7: Age Distriubtion in the Aurora Raw Data - Warfarin Cohort

The number of subjects in warfarin cohort with an age under 18 is 0 . These subjects are excluded from the cohort as the CoumaGen-II's inclusion criteria is subjects with the age of 18 or older.

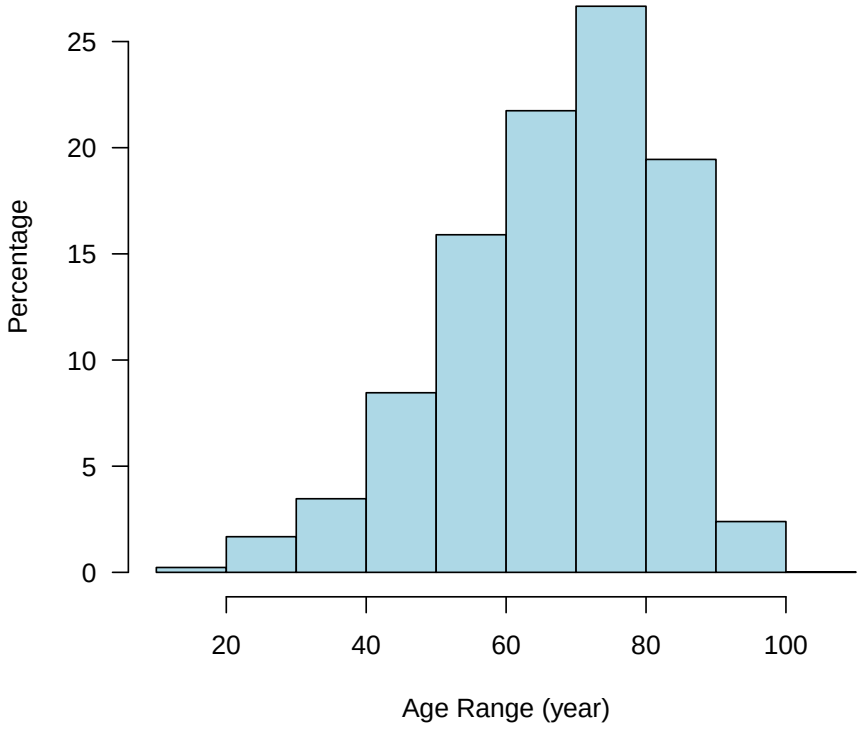

Figure 3: Warfarin Cohort - Age Distribution

5.1.2 Gender

Here is the gender distribution of the warfarin cohort (Table 4) and also that of Milwaukee County and the State of Wisconsin (Table 5) for comparison.

Gender	Percentage
Male	50.83
Female	49.17
Missing	0.00

Table 8: Gender Distriubtion in the Aurora Raw Data - Warfarin Cohort

	MKE(%)	WI(%)
Male	48.30	49.60
Female	51.70	50.40

Table 9: Gender Distriubtion in Milwaukee County (MKE) and Wisconsin (WI).

5.1.3 Race

63.8 percent of warfarin cohort have race information.The following table depicts the race distribution in the warfarin cohort.

Race	Percentage
White	94.15
Black or African American	5.20
Asian	0.45
American Indian or Alaskan Native	0.18
Native Hawaiian/Other Pacific Islander	0.02

Table 10: Warfarin Cohort's Race Distriubtion

5.1.4 Height

In the ARD, each subject has multiple height records. Height has the following attributes: SURROGATE_ID, EFFECTIVE_DAY, HEIGHT, SOURCE_SYSTEM

For the height records, a few steps are taken: First, to identify and exclude height records whose HEIGHT attributes are missing. Second, to identify and exclude duplicates height records. Third, since in the final modeling and simulation study each subject should have one height record and given that each subject should have been almost stable in height during the treatment period at AHC, we take the median of each subject's height records and take it as the height of the subject. Third, the height medians are refined by excluding the ones that here are considered outliers (i.e., median heights over or lower 3 standard deviations). Fourth, to identify and extract the height of warfarin cohort.

```
The average height of the warfarin cohort after excluding the outliers is 66.96 inch.
```

```
[1] 157428
```

```
[1] 11
```