6.S897/HST.956 Machine Learning for Healthcare

Lecture 3: Deep Dive into Clinical Data

Instructors: David Sontag, Peter Szolovits

1 Understanding Clinical Data

To start understanding Clinical Data, we begin with an example from the distribution of heart rates from the MIMIC-III database [EWJJPS+16] (as recorded in Careview), shown in Figure 1. This data involves around 600,000 admissions over a period of 12 years.

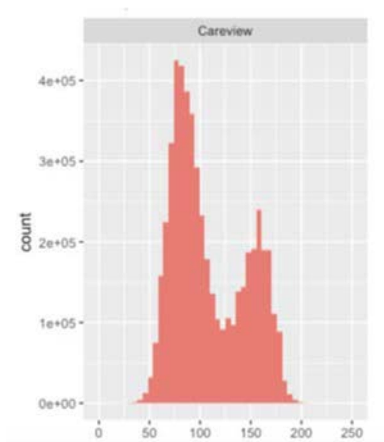


Figure 1: The distribution of heart rates in the Careview medical system.

Unusually for biological data like heart rate, the data is bimodal (we'd typically expect a distribution closer to normal). Why might this be? As it turns out, the hospital that provides this data switched care systems, from Careview to Metavision, and the old and new systems don't record data in exactly the same way. A comparison of the two systems can be seen in Figure 2.

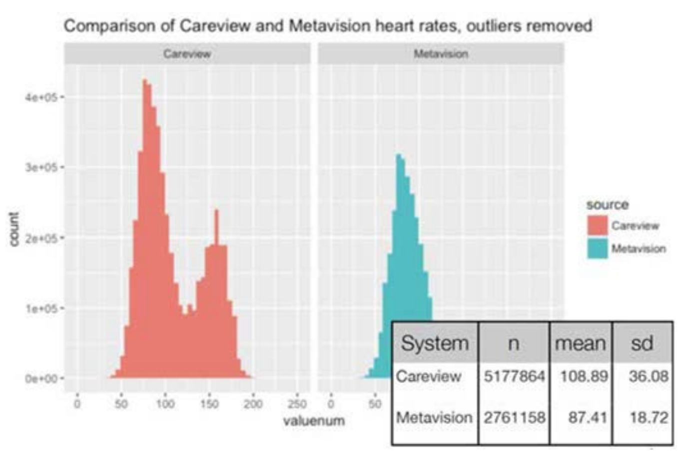


Figure 2: A comparison of the heart rate data in the two different systems.

The data from the Metavision looks closer to normal and is not bimodal, so it looks much more like we would expect. Further investigation reveals two major ways the first system differes from the second. First, under Careview, natal intensive care unit data was added to overall data set, while that data was not included in Metavision. Second, everyone over the age of 90 in the Carevision was listed as 300 years old upon their first visit to the system, in order to protect their identities in compliance with HIPPA regulations. These data are shown in Figure 3.

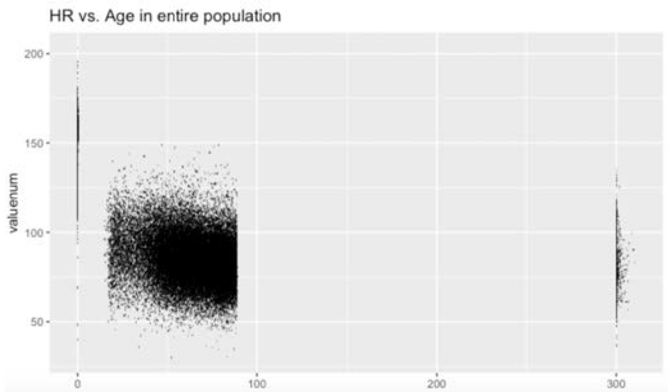


Figure 3: Heart Rate vs Age (Careview).

Once these differences are accounted for, the data actually look quite similar, which one can see in Figure 4. The lesson here is "be careful with data." There are many strange issues with how it's collected and stored that can be confusing without additional information.

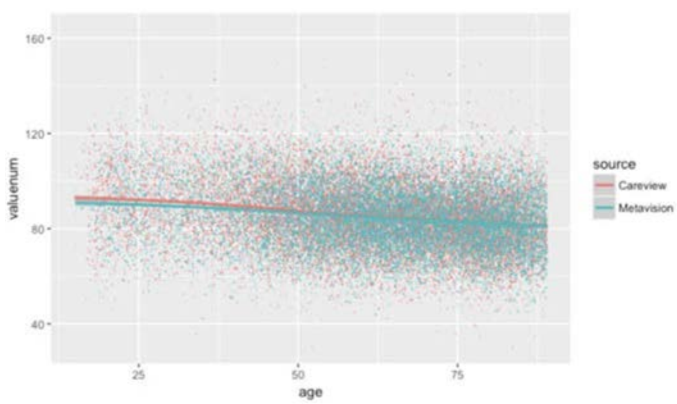


Figure 4: Heart Rate vs Age for adults.

2 Types of Data

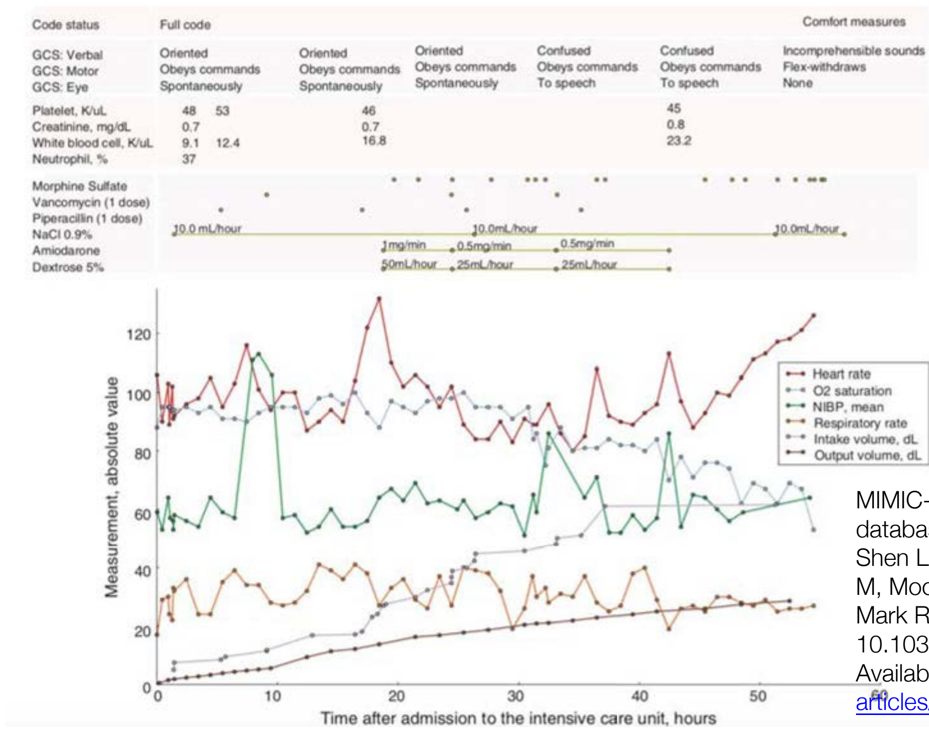
There are many types of health care data we can use, including:

- Demographics – This includes data like age, sex, race, etc.
- Vital signs – These data are basically measurements a nurse would take during a regular check up, like weight, height, blood pressure, etc.

- Medications – These data cover over-the-counter drugs one takes, as well as illegal drugs and alcohol. This is an area that patients could lie about, particularly when illegal drugs are involved. However, with lab results one can get a more accurate picture of the substances a person uses, in a process called "medication reconciliation."
- Lab test results – Components of different bodily fluids, including blood, stool, urine, etc.
- Pathology – This involves qualitative and quantitative examinations of any body tissues, including cell-level measurements such as cell-surface antigens. A rule of thumb is that if something is taken out of you during surgery, it's probably going to Pathology.
- Microbiology – This involves growing organisms, typically from cultures, to test their sensitivity to various antibiotics, at various dilutions, etc.
- Notes – There notes at the end of medical reports, which can be quite long, and contain information like the kind of drugs a patient will be prescribed, whether there was a referring physician who sent the patient in, whether the patient has been advised to seek a specialist, whether the patient will receive in-home care, etc.
- Billing – All the information about what was billed by the hospital. There can be a large amount of information here, because hospitals in general will want to bill for as much as they justifiably can. Includes ICD9/10 codes, procedure codes, etc.
- Administrative data – This involves which service you're on. An example of where this comes up would be that you need cardiology intensive care, but that service is full, and instead you get a bed in pulmonary intensive care. You would still be listed as getting cardiology service, even though your bed is in pulmonary.
- Imaging data – this includes x-rays, ultrasounds, etc.
- Quantified self data – Data that come from wearable devices, including steps walked, elevation change, heart rate, diet, blood sugar, etc.

2.1 Example Chart

An example of the kinds of charts we might deal with is in Figure 5, which is a chart going over the care of a person in an intensive care unit from their admittance to their death.



MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635

Figure 5: An example medical chart.

The top row of the chart covers the amount the patient wants doctors to try to keep them alive if something goes wrong at the time in intensive care. "Full code" means that they want every effort to be made to keep them alive, while "Comfort measures" means that they want doctors to allow them to die if their condition worsens. In the chart, when the patient is admitted their code status is"full code," but as time progresses they change to "comfort measures."

The second section measures the physical capabilities of the patient, such as their ability to speak, their motor control, and their eye movements. While at the beginning of their admission they are in full bodily control, their condition deteriorates over their time in intensive care.

The third section covers fluid measurements over the course of the visit, such as platelet count.

The fourth section covers various medications that were administered to the patient over the course of their admission, including the doses involved.

Finally, a chart covers various vital measurement over the course of the visit, such as heart rate and O2 saturation.

2.2 Demographic Comparison

As part of an exploratory analysis, one can plot how demographic variables relate to each other to try to better understand subpopulations of the patients. For example, you could plot age of admission segmented by admission type (elective, emergency, or urgent), as seen in Figure 6. In this case, the distribution of age does not change very much depending on their admission type.

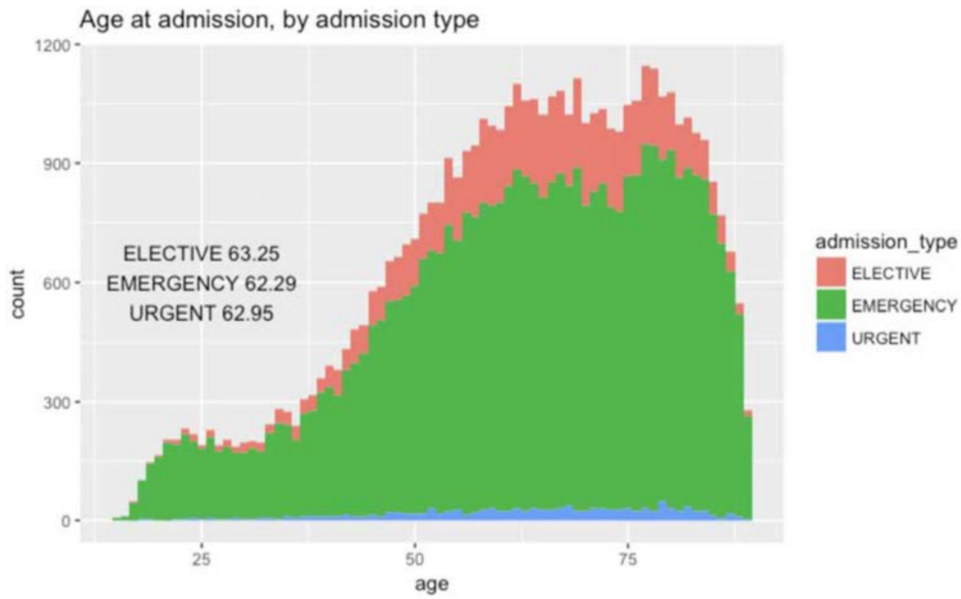


Figure 6: Age of admission, by admission type.

As another example, in Figure 7 we plot age segmented by insurance type. There is a clear change in age distribution here, as self paying customers skew younger, and most people switch to Medicare after age 65.

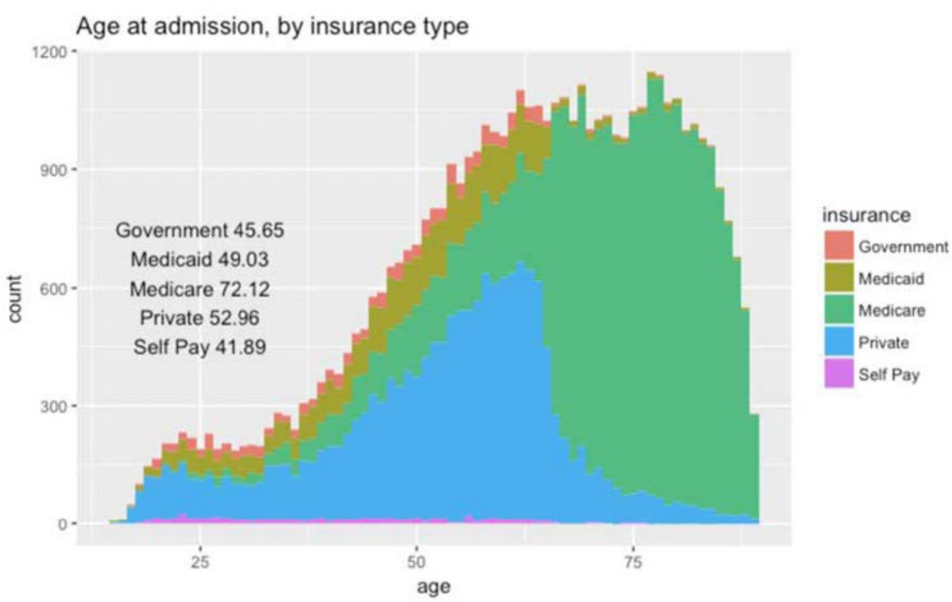


Figure 7: Age of admission, by insurance type.

More examples of these plots can be found in the lecture slides.

One can also investigate how mortality is influenced by demographic information. In Figure 8 we have the results of generalized linear model trained on the health care data to predict mortality, with significant variables indicated by stars.

```
glm(formula = hospital_expire_flag ~ age + ethnicity + marital_status +
    language, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1146  -0.4583  -0.3812  -0.3054   2.8384

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -3.107213   0.651502  -4.769 1.85e-06 ***
age                             0.031763   0.001774  17.901  < 2e-16 ***
ethnicityHISPANIC OR LATINO    -0.013091   0.196425  -0.067 0.946863
ethnicityOTHER                 -0.016074   0.186942  -0.086 0.931477
ethnicityUNABLE TO OBTAIN       0.803709   0.151518   5.304 1.13e-07 ***
ethnicityUNKNOWN/NOT SPECIFIED  0.562160   0.159312   3.529 0.000418 ***
ethnicityWHITE                  0.041665   0.079084   0.527 0.598298
marital_statusMARRIED          -0.009904   0.088537  -0.112 0.910929
marital_statusSEPARATED         0.224446   0.213855   1.050 0.293935
marital_statusSINGLE            0.009709   0.094831   0.102 0.918449
marital_statusWIDOWED          -0.113735   0.102765  -1.107 0.268403
languageENGL                   -1.487467   0.630198  -2.360 0.018259 *
languagePTUN                   -0.754769   0.640661  -1.178 0.238753
languageRUSS                   -1.210058   0.642498  -1.883 0.059651 .
languageSPAN                   -1.311704   0.657075  -1.996 0.045904 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15330  on 27223  degrees of freedom
Residual deviance: 14792  on 27209  degrees of freedom
  (17028 observations deleted due to missingness)
AIC: 14822
```

Figure 8: Age of admission, by insurance type.

Age being significant makes sense, as the older someone is the more likely they are to die. The statistically significant ethnicity variables are those that indicate ethnicity information is missing for some reason, which is difficult to explain – the information is missing frequently enough that it is unlikely it's missing because people die before being able to identify their ethnicity. The remaining significant variables are knowing English or Spanish, which indicates being able to communicate with doctors more easily leads to lower mortality.

3 Different Medical Standards

There are a wide variety of medical standards, for everything from prescriptions given to procedures and more. For example, in Figure 9, we see two patients with the same diagnosis given very different treatments.

SUBJECT_ID	57139	57139
HADM_ID	155470	155470
ICUSTAY_ID	NA	NA
STARTDATE	2185-12-07	2185-12-07
ENDDATE	2185-12-07	2185-12-23
DRUG_TYPE	MAIN	MAIN
DRUG	Acetaminophen	Clobetasol Propionate 0.05%Cream
DRUG_NAME_POE	Acetaminophen	Clobetasol Propionate 0.05%Cream
DRUG_NAME_GENERIC	Acetaminophen	Clobetasol Propionate 0.05%Cream
FORMULARY_DRUG_CD	ACET325	CLOB.05C30
GSN	4489	7634
NDC	182844789	472040030
PROD_STRENGTH	325mg Tablet	30gm Tube
DOSE_VAL_RX	325-650	1
DOSE_UNIT_RX	mg	Appl
FORM_VAL_DISP	1-2	0.01
FORM_UNIT_DISP	TAB	TUBE
ROUTE	PO	TP

Figure 9: Two different treatments for the same diagnosis.

One can also identify different medical standards by looking at the most common prescriptions in the database, as seen in Figure 9. For example, there are two different rows containing D5W, one with an NDC code and one without, along with several other examples of prescriptions without their NDC codes. Thus, even though ways of recording prescriptions like NDC codes are standardized across the US, the ways hospitals report what they prescribe are not necessarily standardized.

Most Common Prescriptions

	NDC Code	count
Iso-Osmotic Dextrose	0	86935
Sodium Chloride 0.9% Flush	0	83392
Insulin	0	81356
SW	0	72458
Magnesium Sulfate	409672924	55211
D5W	0	54938
Furosemide	517570425	53073
Potassium Chloride	338070341	47968
D5W	338001702	43038
LR	338011704	35407
Vancomycin	338355248	34741
0.9% Sodium Chloride	338004904	34682
Potassium Chloride	456066270	32533
Heparin	63323026201	31413
NS	338004902	30815

Figure 10: Most common prescriptions.

3.1 Different Medication Coding Systems

There are a large number of medication coding systems, including:

- National Drug Code (NDC) – A 10 number identification code for a drug where the first four numbers identify who produced it, the next four the form of the drug, and the last two the number of doses. This coding system has the difficulty that it has run out of numbers for both the drug producers and the form of the drug, and attempts to expand it have not been applied systematically.
- MedDRA – An identification system created by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. It isn't compatible with NDC codes.
- CPT Codes – The codes in the range 90281- 99607 give a variety of medicine codes.
- 2019 Healthcare Common Procedure Coding System (HCPCS) – Used by Medicare and Medicaid patients.
- Commercial Coding Systems – While often redundant, Medi-Span and First Data Bank also have created medication codes.

3.2 Different Procedure Coding Systems

There are also a variety of different codes that identify medical procedures patients receive, including:

- ICD 9/10 procedure codes
- CPT codes
- MV codes

CPT codes in particular are highly detailed – example CPT categories can be seen in Figure 11. Note that each category has multiple codes associated with it.

Medicine	90281-90399	Immune globulins, serum or recombinant prods
Medicine	90465-90474	Immunization administration for vaccines/toxoids
Medicine	90476-90749	Vaccines, toxoids
Medicine	90801-90899	Psychiatry
Medicine	90901-90911	Biofeedback
Medicine	90918-90925	End-Stage Renal Disease Services (deleted codes)
Medicine	90935-90999	Dialysis
Medicine	91000-91299	Gastroenterology
Medicine	92002-92499	Ophthalmology
Medicine	92502-92700	Special otorhinolaryngologic services
Medicine	92950-93799	Cardiovascular
Medicine	93875-93990	Noninvasive vascular diagnostic studies
Medicine	94002-94799	Pulmonary
Medicine	95004-95199	Allergy and clinical immunology
Medicine	95250-95251	Endocrinology
Medicine	95803-96020	Neurology and neuromuscular procedures
Medicine	96101-96125	Central nervous system assessments/tests (neuro-cogn
Medicine	96150-96155	Health and behavior assessment/intervention
Medicine	96360-96549	Hydration, therapeutic, prophylactic, diagnostic inj
Medicine	96567-96571	Photodynamic therapy
Medicine	96900-96999	Special dermatological procedures
Medicine	97001-97799	Physical medicine and rehabilitation
Medicine	97802-97804	Medical nutrition therapy
Medicine	97810-97814	Acupuncture
Medicine	98925-98929	Osteopathic manipulative treatment
Medicine	98940-98943	Chiropractic manipulative treatment
Medicine	98960-98962	Education and training for patient self-management
Medicine	98966-98969	Non-face-to-face nonphysician services
Medicine	99000-99091	Special services, procedures and reports
Medicine	99170-99199	Other services and procedures
Medicine	99500-99602	Home health procedures/services
Medicine	99605-99607	Medication therapy management services

Figure 11: CPT code categories.

4 Lab Reports

The most commonly used coding system for reporting lab results is Logical Observation Identifiers Names and Codes (LOINC). It has a hierarchical structure where every lab test has a different code, but similar results are connected by categories. An example of a lab report can be seen in Figure 12. Note that the times of the tests have been de-identified, and some qualitative values like atypical lymphocytes do not have flags.

subj	hadm	item	time	value	units	flag	label	fluid	categ	loinc
2	163353	51143	2138-07-17 20:48:00	0.00	%	NA	Atypical Lymphocytes	Blood	Hem	733-6
2	163353	51144	2138-07-17 20:48:00	0.00	%	NA	Bands	Blood	Hem	763-3
2	163353	51146	2138-07-17 20:48:00	0.00	%	NA	Basophils	Blood	Hem	704-7
2	163353	51200	2138-07-17 20:48:00	0.00	%	NA	Eosinophils	Blood	Hem	711-2
2	163353	51221	2138-07-17 20:48:00	0.00	%	abnormal	Hematocrit	Blood	Hem	4544-3
2	163353	51222	2138-07-17 20:48:00	0.00	g/dL	abnormal	Hemoglobin	Blood	Hem	718-7
2	163353	51244	2138-07-17 20:48:00	0.00	%	NA	Lymphocytes	Blood	Hem	731-0
2	163353	51248	2138-07-17 20:48:00	0.00	pg	abnormal	MCH	Blood	Hem	785-6
2	163353	51249	2138-07-17 20:48:00	0.00	%	abnormal	MCHC	Blood	Hem	786-4
2	163353	51250	2138-07-17 20:48:00	0.00	fL	abnormal	MCV	Blood	Hem	787-2
2	163353	51251	2138-07-17 20:48:00	0.00	%	NA	Metamyelocytes	Blood	Hem	28541-1
2	163353	51254	2138-07-17 20:48:00	0.00	%	NA	Monocytes	Blood	Hem	742-7
2	163353	51255	2138-07-17 20:48:00	0.00	%	NA	Myelocytes	Blood	Hem	26498-6
2	163353	51256	2138-07-17 20:48:00	100.00	%	NA	Neutrophils	Blood	Hem	761-7
2	163353	51265	2138-07-17 20:48:00	5.00	K/uL	abnormal	Platelet Count	Blood	Hem	777-3

Figure 12: Lab results for a patient.

5 Chart Events

Chart Events capture a variety of vital sign features. Figure 13 contains some examples of such events. Some, like heart rate, appear twice. This could be an indication of two systems being combined to get these codes, which is something to watch out for.

itemid	n	label	category	units	param_type
211	5180809	Heart Rate	NA	NA	NA
742	3464326	calprevflg	NA	NA	NA
646	3418917	SpO2	NA	NA	NA
618	3386719	Respiratory Rate	NA	NA	NA
212	3303151	Heart Rhythm	NA	NA	NA
161	3236350	Ectopy Type	NA	NA	NA
128	3216866	Code Status	NA	NA	NA
550	3205052	Precautions	NA	NA	NA
1125	2955851	Service Type	NA	NA	NA
220045	2762225	Heart Rate	Routine Vital Signs	bpm	Numeric
220210	2737105	Respiratory Rate	Respiratory	insp/min	Numeric
220277	2671816	O2 saturation pulseoxymetry	Respiratory	%	Numeric
159	2544519	Ectopy Frequency	NA	NA	NA
1484	2261065	Risk for Falls	NA	NA	NA
51	2096678	Arterial BP [Systolic]	NA	NA	NA
8368	2085994	Arterial BP [Diastolic]	NA	NA	NA

Figure 13: Chart events for a patient.

There are also charts that capture patient outputs, like urine or stool samples. An example can be seen in Figure 14.

itemid	n	label	category	units
40055	1917421	Urine Out Foley	NA	NA
226559	1186717	Foley	Output	mL
40076	152716	Chest Tubes CTICU CT 1	NA	NA
43175	108982	Urine .	NA	NA
40054	81828	Stool Out Stool	NA	NA
226588	81128	Chest Tube #1	Output	mL
40069	69467	Urine Out Void	NA	NA

Figure 14: Outputs for a patient.

There are also tables for patient inputs, which include medications provided to them. Two examples of this type of table, one for CareVue and one for MetaVision), can be seen in Figure 15 and Figure 16 respectively.

itemid	n	label
30013	2557507	D5W
30018	2392372	.9% Normal Saline
30131	924614	Propofol
30045	825758	Insulin
30025	813242	Heparin
30118	780555	Fentanyl
30128	554582	Neosynephrine-k
30124	505509	Midazolam
30120	476971	Levophed-k
30140	373023	N/A

Figure 15: Inputs for a patient (CareVue).

itemid	n	label	category	unit	param_type
225158	527855	NaCl 0.9%	Fluids/Intake	mL	Solution
220949	406345	Dextrose 5%	Fluids/Intake	mL	Solution
225943	246312	Solution	Fluids/Intake	mL	Solution
222168	178819	Propofol	Medications	mg	Solution
226452	135438	PO Intake	Fluids/Intake	mL	Solution
223258	119668	Insulin -	Medications	units	Solution
225799	97629	Gastric Meds	Fluids/Intake	mL	Solution
221749	93571	Phenylephrine	Medications	mg	Solution
221906	89697	Norepinephrine	Medications	mg	Solution
221744	86340	Fentanyl	Medications	mg	Solution

Figure 16: Inputs for a patient (MetaVision).

6 Using Medical Process Measures to Make Predictions

The required reading for this class, *Biases in electronic health record data due to processes within the healthcare system* [AKW18], showed that for many lab results, process measures of the data (such as the time a lab result was taken) are more important than actual values in predicting outcomes. While these results cannot be replicated exactly with the MIMIC III database, there are some proxies we can use to try to get similar results, such as white blood cell (WBC) counts. Looking at the fractions of abnormal white blood cell counts per hour, for example, matches the paper's findings that tests taken in the early morning such as 4:00 am are connected to a person being unhealthy. The graph of this relationship can be seen in Figure 17.

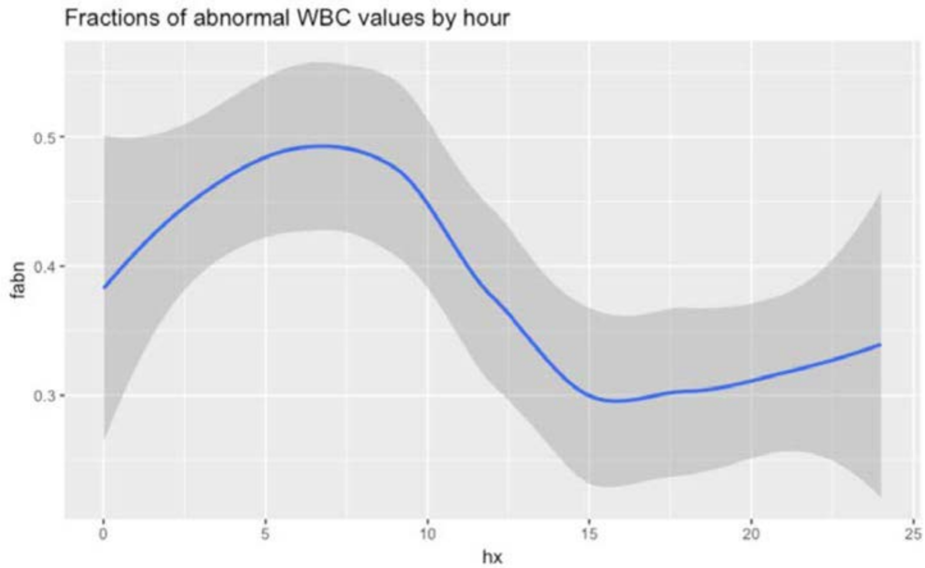


Figure 17: Proportion of abnormal WBC measurements per hour.

One can also build a regression model to predict mortality from number of WBC measurements and number abnormal WBC measurements per hour. This can be found in Figure 18.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8045  -1.0958  -0.5012   1.1245   2.3401

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.04321    0.11487   0.376 0.706758
H0            0.75871    0.88579   0.857 0.391700
H1            0.45657    0.76061   0.600 0.548333
H2            0.39502    0.65687   0.601 0.547597
H3           15.46281  413.03082   0.037 0.970136
H4            0.87956    0.90070   0.977 0.328804
H5            0.19184    0.92995   0.206 0.836562
H6            0.43533    0.65352   0.666 0.505330
H7            0.05389    0.40893   0.132 0.895147
H8            1.36632    0.47436   2.880 0.003972 **
H9            0.07131    0.24685   0.289 0.772685
H10           0.02999    0.16509   0.182 0.855845
H11          -1.03418    0.32225  -3.209 0.001331 **
H12           0.15791    0.21427   0.737 0.461133
H13          -0.39467    0.31470  -1.254 0.209803
H14          -0.19412    0.18526  -1.048 0.294726
H15          -0.42509    0.15821  -2.687 0.007212 **
H16           0.24009    0.12191   1.969 0.048900 *
H17          -0.10166    0.15254  -0.666 0.505139
H18          -0.10116    0.18002  -0.562 0.574149
H19          -0.23376    0.24193  -0.966 0.333919
H20          -0.12929    0.18466  -0.700 0.483827
H21          -0.79920    0.27154  -2.943 0.003248 **
H22          -0.56242    0.36065  -1.559 0.118893
H23          -0.45735    0.47557  -0.962 0.336199
H24           0.08659    0.71026   0.122 0.902962
HA0          -1.78217    1.32944  -1.341 0.180071
HA1          -0.80485    1.28716  -0.625 0.531782
HA2          -1.39389    1.36913  -1.018 0.308639
HA3         -15.69112  413.03210  -0.038 0.969696
HA4          -0.91247    1.21520  -0.751 0.452723
HA5          -0.32100    1.38380  -0.232 0.816564
HA6          -1.32274    1.04715  -1.263 0.206524
HA7          -0.71769    0.93684  -0.766 0.443632
HA8          -1.71813    0.66992  -2.565 0.010327 *
HA9          -0.67054    0.51100  -1.312 0.189450
HA10         -0.19831    0.45897  -0.432 0.665693
HA11          1.72924    0.52482   3.295 0.000984 ***
HA12          0.03971    0.59225   0.067 0.946540
HA13          0.94444    0.62952   1.500 0.133550
HA14          0.22134    0.45705   0.484 0.628188
HA15          1.25147    0.44487   2.813 0.004906 **
HA16          0.04059    0.39246   0.103 0.917633
HA17          0.18535    0.46846   0.396 0.692352
HA18          0.49504    0.44025   1.124 0.260823
HA19         -0.02478    0.45548  -0.054 0.956612
HA20          0.41568    0.53548   0.776 0.437594
HA21          1.60231    0.60935   2.630 0.008550 **
HA22          0.52832    0.56629   0.933 0.350848
HA23          0.92591    0.88156   1.050 0.293580
HA24          0.67132    1.68820   0.398 0.690887
---
Signif. codes:  0 â€˜***â€™ 0.001 â€˜**â€™ 0.01 â€˜*â€™ 0.05
```

Figure 18: Results for using regression to predict mortality from number of WBC measurements and number abnormal WBC measurements per hour.

While there are some significant hours here, the fact that only some hours are significant and not others is not what one would expect based on [AKW18]. For example, if 8:00 am were a significant time to get a lab test done, one would think 7:00 am and 9:00 am would also be significant times, but in this regression that is not the case. Thus, it seems like there is some noise in the data causing the times to appear insignificant.

We can use the MIMIC data to confirm that lab result values do vary by time of day, as the tables in Figure 19 show. This is one example of many tables in the lecture slides showing how lab tests results change over the day depending on the time it is performed. While there are several possible explanations for this, such as diurnal human body changes or changes in care throughout the day, these results do align with the results from [AKW18] that the time a test is taken provides valuable information alongside the test results.

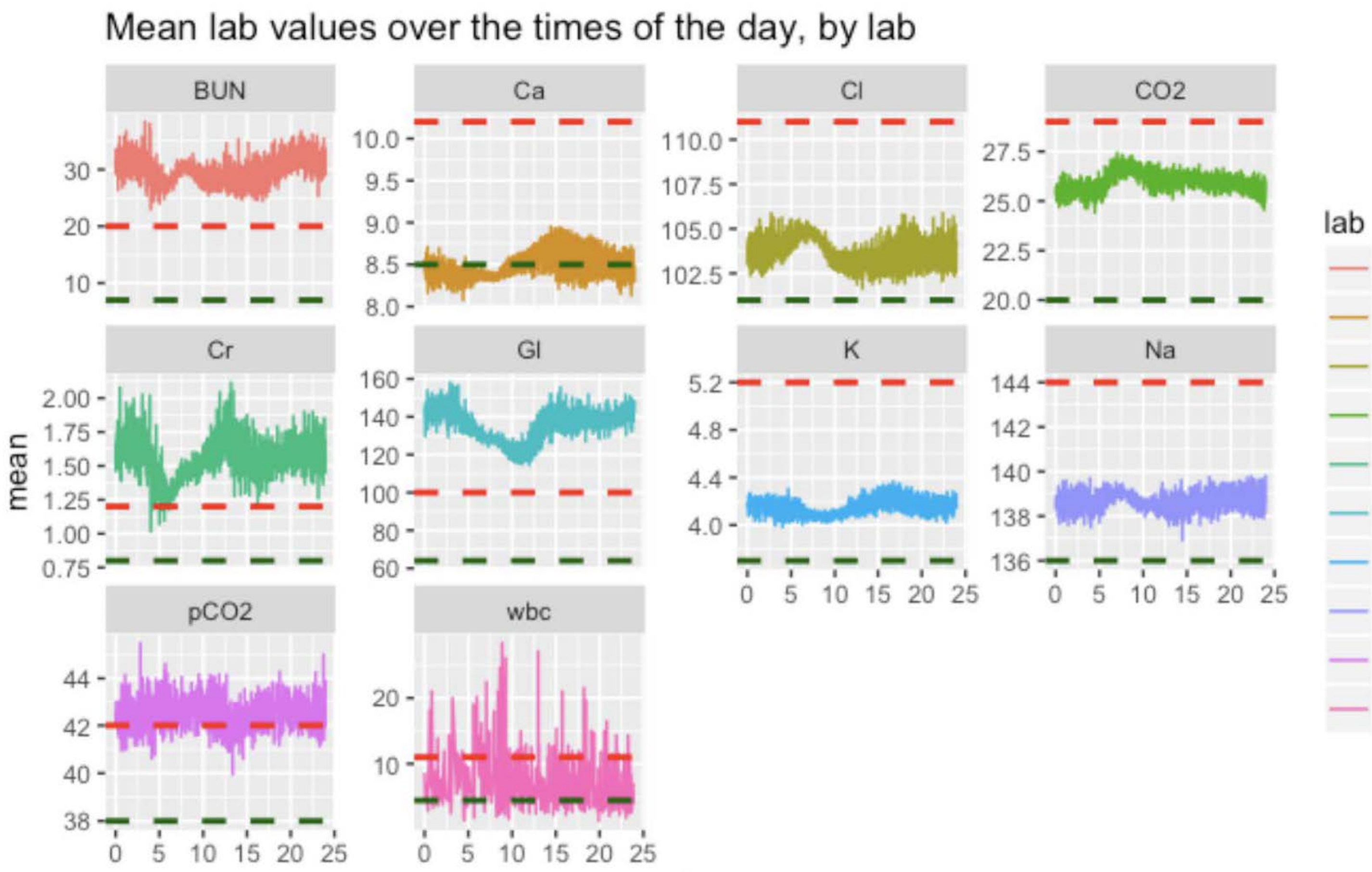


Figure 19: Mean lab result values plotted against time for several lab tests.

7 Clinical Notes in MIMIC

There are a wide variety of types of clinical notes – counts for the number of clinical notes taken by type of professional or visit in the data set can be found in Figure 20.

Nursing/other	822497
Radiology	522279
Nursing	223556
ECG	209051
Physician	141624
Discharge summary	59652
Echo	45794
Respiratory	31739
Nutrition	9418
General	8301
Rehab Services	5431
Social Work	2670
Case Management	967
Pharmacy	103
Consult	98

Figure 20: Different counts of the number of clinical notes in the MIMIC database by profession or reason for visit.

These notes can be very long. The empirical counts of the lengths of these notes taken by type of profession can be found in Figure 21 – average counts of 1000 words or more are not uncommon.

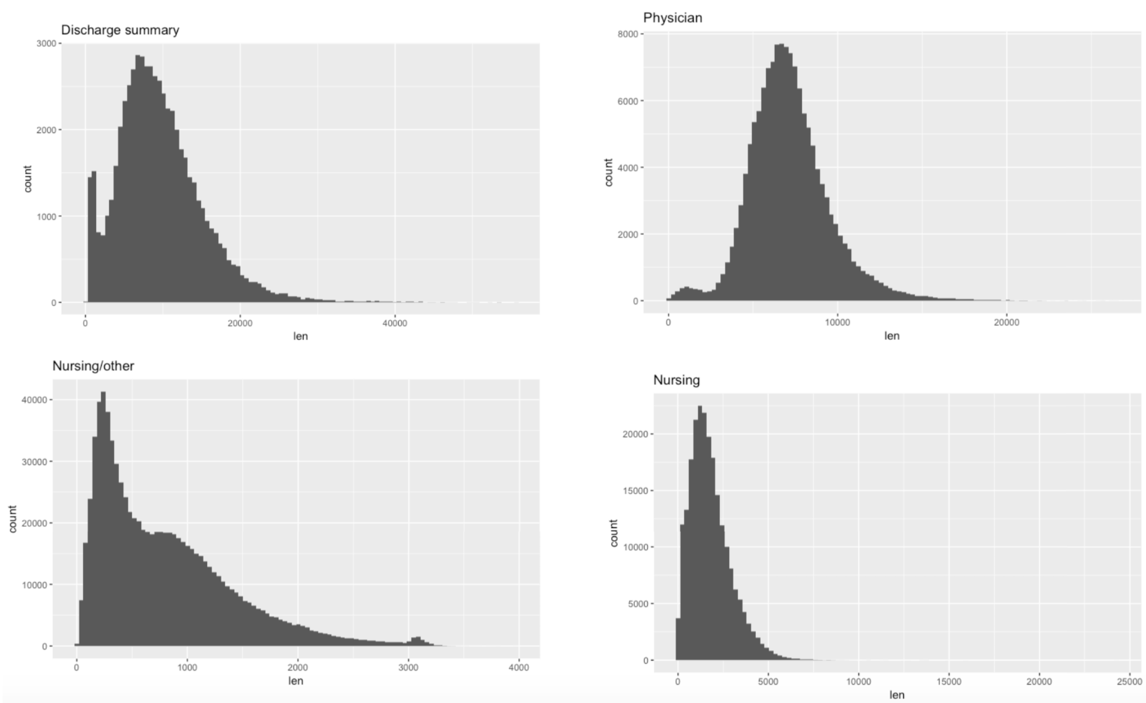


Figure 21: The distribution of the lengths of the clinical notes by profession or reason for visit.

An example nursing note can be found in the lecture slides.

8 Data Encoding Standards

OHDSI is the standard data encoding method. It is often used with Fast Healthcare Interoperability Resources (FHIR), which allows hospitals to share healthcare information electronically. The goal of FHIR is to provide the minimum amount of information a doctor needs to know to start treating a patient. Figure 22 shows an example of the form of healthcare information shared – various applications make the form easier for humans to parse.

```xml
<Patient xmlns="http://hl7.org/fhir">
  <id value="glossy"/>
  <meta>
    <lastUpdated value="2014-11-13T11:41:00+11:00"/>
  </meta>
  <text>
    <status value="generated"/>
    <div xmlns="http://www.w3.org/1999/xhtml">
      <p>Henry Levin the 7th</p>
      <p>MRN: 123456. Male, 24-Sept 1932</p>
    </div>
  </text>
  <extension url="http://example.org/StructureDefinition/trials">
    <valueCode value="renal"/>
  </extension>
  <identifier>
    <use value="usual"/>
    <type>
      <coding>
        <system value="http://hl7.org/fhir/v2/0203"/>
        <code value="MR"/>
      </coding>
    </type>
    <system value="http://www.goodhealth.org/identifiers/mrn"/>
    <value value="123456"/>
  </identifier>
  <active value="true"/>
  <name>
    <family value="Levin"/>
    <given value="Henry"/>
    <suffix value="The 7th"/>
  </name>
  <gender value="male"/>
  <birthDate value="1932-09-24"/>
  <careProvider>
    <reference value="Organization/2"/>
    <display value="Good Health Clinic"/>
  </careProvider>
</Patient>
```


Figure 22: The distribution of the lengths of the clinical notes by profession or reason for visit.

9 Resources for Various Terminologies

All of the terminology standards, such as LOINC, ICD9/10, etc., are gathered in the UMLS Metathesaurus at https://uts.nlm.nih.gov/home.html.

10 Key Takeaways

- "Know your data"
- Harmonising all the different types of data is difficult and time consuming
- For some areas, standards don't exist at all

References

[AKW18] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361, 2018.

[EWJJPS+16] Alistair Edward William Johnson, Tom Joseph Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Edward Moody, Peter Szolovits, Leo Anthony G. Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 05 2016.