Preprocessing Electronic Health Records for Analysis-Ready Data in an Asthma Cohort

Kimberly Lactaoen, MS

Thursday, June 12, 2025

Penn

EHR data

- EHR data captures extensive patient data for large-scale studies
- However, it is complex, subject to inaccuracies, and prone to missingness which may not represent a patient's true health status[1]
- Demonstrate specific challenges when preparing data for analysis
  - Tidyverse (dplyr, lubridate, stringr, and tidyr)

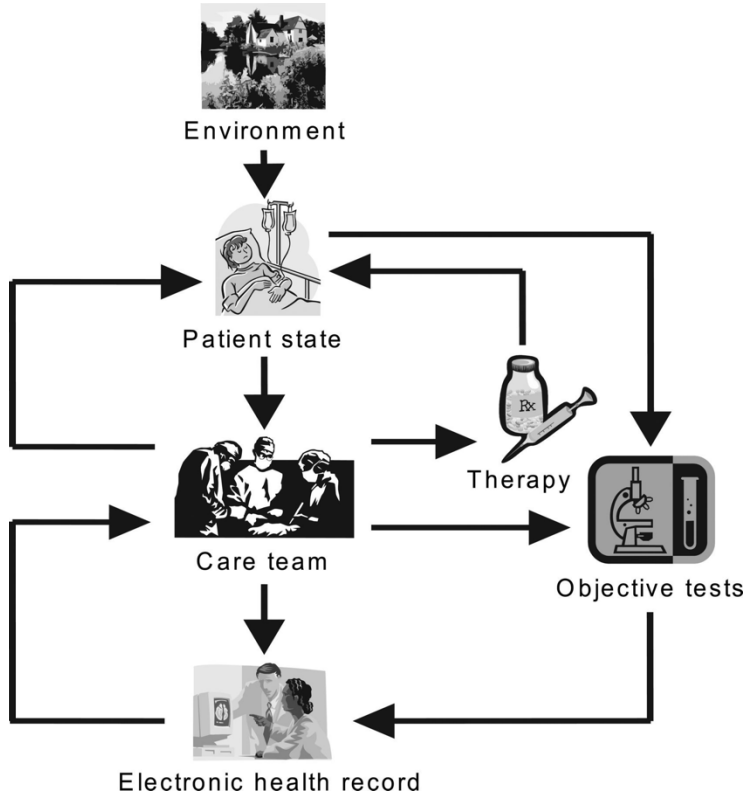

Figure 1. Feedback loops in the electronic health record.

J Am Med Inform Assoc, Volume 20, Issue 1, January 2013, Pages 117–121, https://doi.org/10.1136/amiajnl-2012-001145

[1]Hubbard RA, Lou C, Himes BE: The Effective Sample Size of EHR-Derived Cohorts Under Biased Sampling. Modern Statistical Methods for Health Research. Zhao Y, Chen D (eds.). Springer International Publishing, 2021

EHR data collection

- Data collected from the Clarity database from PennChart (EPIC)

- Request data relevant to asthma research
  - Demographics
  - Diagnosis codes
  - Medications
  - Laboratory test results

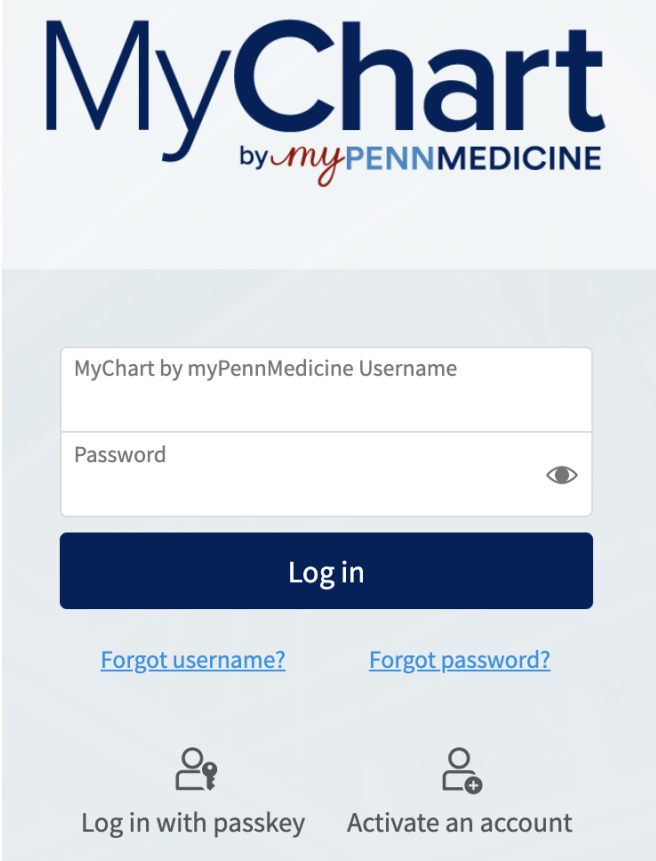

Figure 2. Screenshot of PennChart login page.

Demographics

- **Challenge:** understanding how data was collected
- **Background info:** demographic information is presented by encounter level
- **Resolution:** communication with data analyst
  - *Sex*, *Race*, *Ethnicity*, and *Smoking* status were captured from Patient Table
    - Used the most recent entry
  - *Insurance* and *BMI* variables change over time

Patient_ID	Enc_ID	Date	Sex	Race	BMI	Ethnicity	Smoking	Insurance
1	kdl1	2023-01-05	F	White	28	non-Hispanic/Latino	Ever	Private
2	kdl2	2025-10-13	M	Black	31	non-Hispanic/Latino	Never	Medicare
1	kdl3	2025-03-22	F	White	29	non-Hispanic/Latino	Ever	Private
1	kdl4	2025-05-30	F	White	29	non-Hispanic/Latino	Ever	Unknown
3	kdl5	2025-06-12	M	White	31	non-Hispanic/Latino	Unknown	Private
4	kdl6	2025-04-27	M	Asian	32	Hispanic/Latino	Unknown	Medicaid

Figure 3. Example of demographic characteristics from EHR data.

Demographics

```r
sex <- enc |>
  group_by(patient_ID) |>
  summarise(sex_combined = paste(unique(sex), collapse = ", ")) |>
  ungroup() ##same
table(sex$sex_combined, useNA = "ifany")

race <- enc |>
  group_by(patient_ID) |>
  summarise(race_combined = paste(unique(race), collapse = ", ")) |>
  ungroup() ##same
table(race$race_combined, useNA = "ifany")

ethnicity <- enc |>
  group_by(patient_ID) |>
  summarise(ethnicty_combined = paste(unique(ethnicity), collapse = ", ")) |>
  ungroup() ##same
table(ethnicity$ethnicty_combined, useNA = "ifany")

smoking <- enc |>
  group_by(patient_ID) |>
  summarise(smoking_combined = paste(unique(smoking), collapse = ", ")) |>
  ungroup() ##same
table(smoking$smoking_combined, useNA = "ifany")

bmi <- enc |>
  group_by(patient_ID) |>
  summarise(BMI_combined = paste(unique(BMI), collapse = ", ")) |>
  ungroup() ##different
table(bmi$BMI_combined, useNA = "ifany")

insurance <- enc |>
  group_by(patient_ID) |>
  summarise(insurance_combined = paste(unique(insurance), collapse = ", ")) |>
  ungroup() ##different
table(insurance$insurance_combined, useNA = "ifany")
```

```r
Table1 <- enc |>
  group_by(patient_ID) |>
  mutate(Date = as_date(Date)) |>
  filter(Date == (max(Date))) |>
  filter(row_number()==1) |>
  select(patient_ID, sex, race, ethnicity, smoking) |>
  distinct()
```

Figure 4. Example of shortening R script after understanding how the data was collecting, eliminating much of the exploratory analysis.

Demographics

- **Challenge:** conflicting patient information

- **Background info:** The *Race* variable has Hispanic/Latino-Black(HLB) and Hispanic/Latino-White(HLW) values where some patients with these values also have non-Hispanic/Latino as their *Ethnicity*

- **Resolution:** those that were denoted HLB or HLW in *Race* AND Hispanic/Latino in *Ethnicity,* were recoded as Black or White, respectively, in *Race*. Those with conflicting information were recoded as Unknown.

Patient_ID	Enc_ID	Race	Ethnicity
5	kdl7	HLB	non-Hispanic/Latino
6	kdl8	White	non-Hispanic/Latino
7	kdl9	Black	non-Hispanic/Latino
8	kdl10	HLW	Hispanic/Latino

Figure 5. Example of Race and Ethnicity variables from EHR data.

Demographics

```r
enc_update <- enc |>
  mutate(Race = case_when(patient_ID %in% Race_Ethnicity_conflict$patient_ID ~ "Unknown",
                          .default = as.character(Race))) |>
  mutate(Ethnicity = case_when(patient_ID %in% Race_Ethnicity_conflict$patient_ID ~ "Unknown",
                          .default = as.character(Ethnicity))) |>
  mutate(Race = case_when(patient_ID %in%
                            no_conflict_HLW$patient_ID ~ "White",
                          .default = as.character(Race))) |>
  mutate(Race = case_when(patient_ID %in%
                            no_conflict_HLB$patient_ID ~ "Black",
                          .default = as.character(Race)))
```

Figure 6. Example of using case_when function from dplyr package.

Diagnostic codes

- **Challenge:** Diagnostic code descriptions are not always consistent with diagnostic code

- **Background info:** EHR data has diagnostic codes and diagnostic code descriptions. As an example, J45.2 is the ICD-10 code for Mild intermittent asthma, uncomplicated

- **Resolution:** link diagnostic code information directly from government-based ICD-10 files

Patient_ID	Enc_ID	ICD-10	Description
5	kdl7	J45.2	Mild intermittent asthma without complication
6	kdl8	J45.2	Mild intermittent asthma
7	kdl9	J45.2	Mild intermittent asthma, uncomplicated

Figure 7. Example of ICD-10 codes and their descriptions in EHR data.

Diagnostic codes

```r
dx <-read_delim("/project/diagnosis.txt") |>
  mutate(icd_code =
           str_replace_all(CODE, "\\.", ""))

lines <- read_lines("/project/icd10cm_order_2025.txt")

parts_df <- tibble(
  icd_code = map_chr(str_split(lines, "\\s+"), ~ .x[2]),
  icd_definition = str_sub(lines, 78)
)

dx_updated <- left_join(dx, parts_df, by = "icd_code") |>
  select(patient_ID, Date, icd_code, icd_definition)
```

Figure 8. Example of using functions from readr, stringr, and dplyr packages.

Diagnostic codes

```r
dx_bypatient <- dx_updated |>
  select(patient_ID, icd_code) |>
  distinct() |>
  mutate(value = 1) |>
  pivot_wider(
    names_from = icd_code,
    values_from = value,
    values_fill = list(value = 0))
```

patient_ID	icd_code
1	J45.2
1	J45.3
1	J45.5
2	J45.2
2	D50.9
...	...

patient_ID	J45.2	J45.3	J45.5	D50.9	M54.2	...
1	1	1	1	0	0	...
2	1	0	0	1	0	...
3	1	1	1	1	1	...

Figure 9. Example of using pivot_wider function from tidyr packages to create data frame of diagnostic codes by patient.

Medications

- **Challenge:** selecting medications relevant to asthma treatment
- **Background info:** We searched for therapeutic and pharmacological classes like respiratory agents and corticosteroids and narrowed our search to medications like inhaled corticosteroids (ICS) and oral corticosteroids (OCS). However, not all specific medication names under these categories were likely to treat asthma.
- **Resolution:** Individually reviewed each medication name

NAME	MED_ROUTE
AEROBID INHALATION	INHALATION
ALLERGY RELIEF NASAL	NASAL
ALVESCO 160 MCG/ACT INHALATION AERS	INHALATION
ALVESCO 160 MCG/ACT INHALATION AERS	NULL
ALVESCO 80 MCG/ACT INHALATION AERS	INHALATION
ALVESCO 80 MCG/ACT INHALATION AERS	NULL
ALVESCO 80 MCG/ACT INHALATION AERS	ORAL
ALVESCO INHALATION	INHALATION

Figure 10. Example of medication names and medication routes.

Medications

NAME	MED_ROUTE	Keep
BENRALIZUMAB 30 MG/ML SC SOAJ	SUBCUTANEOUS	yes
BENRALIZUMAB 30 MG/ML SC SOSY	INJECTION	yes
BENRALIZUMAB 30 MG/ML SC SOSY	NULL	yes
BENRALIZUMAB 30 MG/ML SC SOSY	SUBCUTANEOUS	yes
BENRALIZUMAB 30 MG/ML SC SOSY	SUBCUTANEOUS INFUSION	yes
BENRALIZUMAB SC	SUBCUTANEOUS	yes
CINQAIR 100 MG/10ML IV SOLN	INTRAVENOUS	yes
CINQAIR 100 MG/10ML IV SOLN	NULL	yes
CINQAIR IV	INTRAVENOUS	yes
CINQAIR IV	NULL	yes

LAMA_route | biologic_route | ICS_LABA_route | ICS_LABA

```r
med_route <- c("OCS_route", "ICS_route", "SABA_route", "SAMA_route", "LABA_route",
               "LAMA_route", "biologic_route", "ICS_LABA_route",
               "ICS_LABA_LAMA_route", "SABA_SAMA_route", "LABA_LAMA_route")

meds_to_keep <- map_dfr(med_route, ~
  read_excel("/project/unique_med_route.xlsx", sheet = .x) |>
    select(NAME, MED_ROUTE, Keep)
)
```

Figure 11. Example of manually reviewing medication data then using map_dfr function from the purr package to create one data frame of medications that will be used to filter for relevant asthma medications.

Laboratory test results

- **Challenge:** not all facilities report the same units of measurement
- **Background info:** Eosinophil units of measures vary in this codified dataset and some units are not reported
- **Resolution:** Convert units of measures to one unit. For those eosinophil measures without a unit of measure, either remove from data, make best assumptions, or chart review.

```
eosinophils_updated <- eosinophils |>
  mutate(REFERENCE_UNIT = case_when(REFERENCE_UNIT %in% c("THO/uL",
                                                         "K/uL",
                                                         "Thousand/uL",
                                                         "x10E3/uL",
                                                         "Thou/uL",
                                                         "x10(3)/mcl",
                                                         "x10*3/uL",
                                                         "x10E3/ul",
                                                         "k/ul",
                                                         "K/UL",
                                                         "x10E3/u") ~ "10*3/uL",
                                    REFERENCE_UNIT == "cells/ul" ~ "cells/uL",
                                    REFERENCE_UNIT == "K/cu mm" ~ "k/cumm",
                                    .default = REFERENCE_UNIT)) |>
  mutate(ORD_VALUE = as.numeric(ORD_VALUE)) |>
  mutate(ORD_VALUE = case_when(REFERENCE_UNIT == "cells/uL" ~ ORD_VALUE/1000,
                               .default = ORD_VALUE)) |>
  mutate(REFERENCE_UNIT = case_when(REFERENCE_UNIT %in% c("k/cumm",
                                                         "cells/uL") ~ "10*3/uL",
                                    .default = REFERENCE_UNIT)) |>
  rename(eos_value = ORD_VALUE)
```

Figure 12. Example script for converting multiple units of measure to one unite in eosinophil data.

Conclusion

- Understand how the data was collected
  - Example: Patient Table and encounter-level demographic variables
- Be on the lookout for conflicting information
  - Example: Race and Ethnicity variables
- Be aware of unspecified information
  - Example: ICD-10 codes
- Remember to collaborate with experts
  - Example: medication names
- Simplify data when you can
  - Example: eosinophil units of measure

Thank you!

- Blanca Himes, PhD
- Gary Weissman, MD, MS

Email: kimberly.lactaoen@pennmedicine.upenn.edu