Work1：剩余2.4T

Echo 1.7T + ECG 91G + Note 2G + Emergency Dept 1G + hosp&icu 10G + MIMIC III CareVue 3G≈1.8T

Echo：<https://physionet.org/content/mimic-iv-echo/0.1/>

ECG：<https://physionet.org/content/mimic-iv-ecg/1.0/>

Note：<https://physionet.org/content/mimic-iv-note/2.2/>

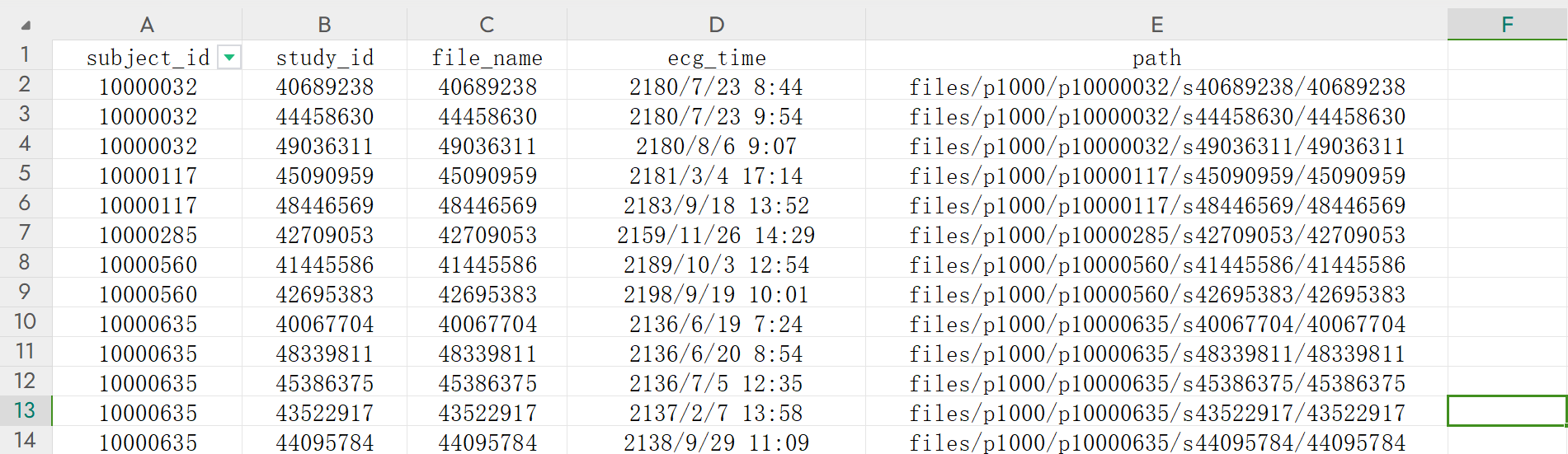
ED：<https://physionet.org/content/mimic-iv-ed/2.2/>

Hosp&icu：<https://physionet.org/content/mimiciv/3.0/>

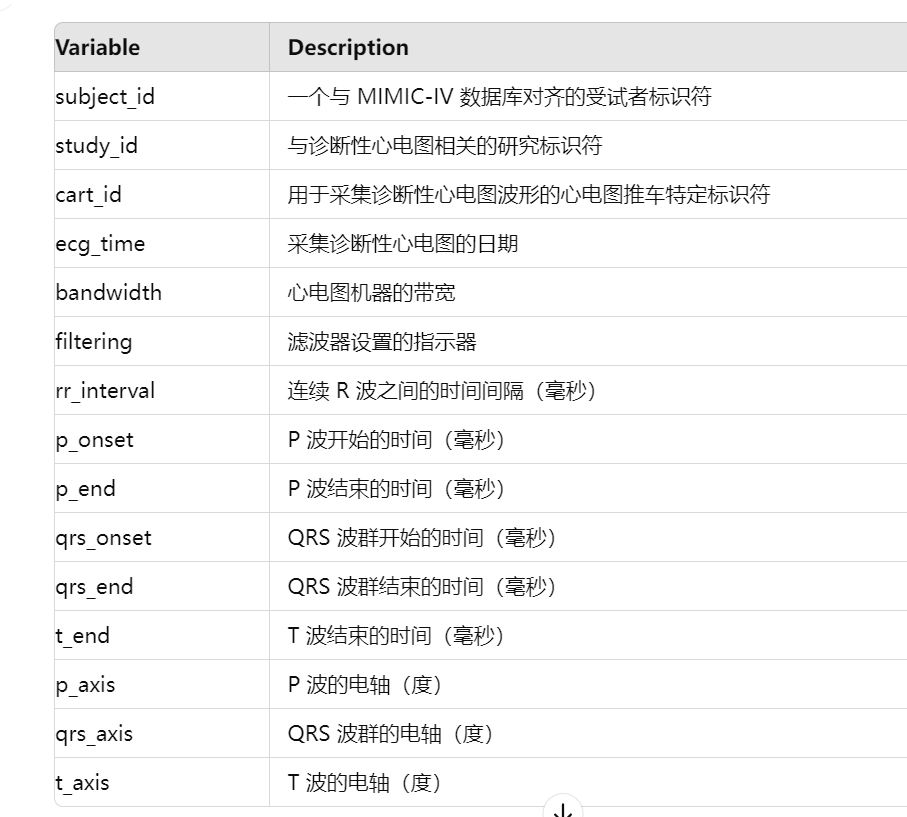
MIMIC III CareVue ：<https://physionet.org/content/mimic3-carevue/1.4/>

# ECG database：

1、record\_list.csv 文件包含每个 WFDB 记录的文件名和路径。它还提供相应的受试者 ID 和研究 ID。受试者 ID 可用于将 MIMIC-IV-ECG 中的受试者链接到 MIMIC-IV 临床数据库中的其他模块。

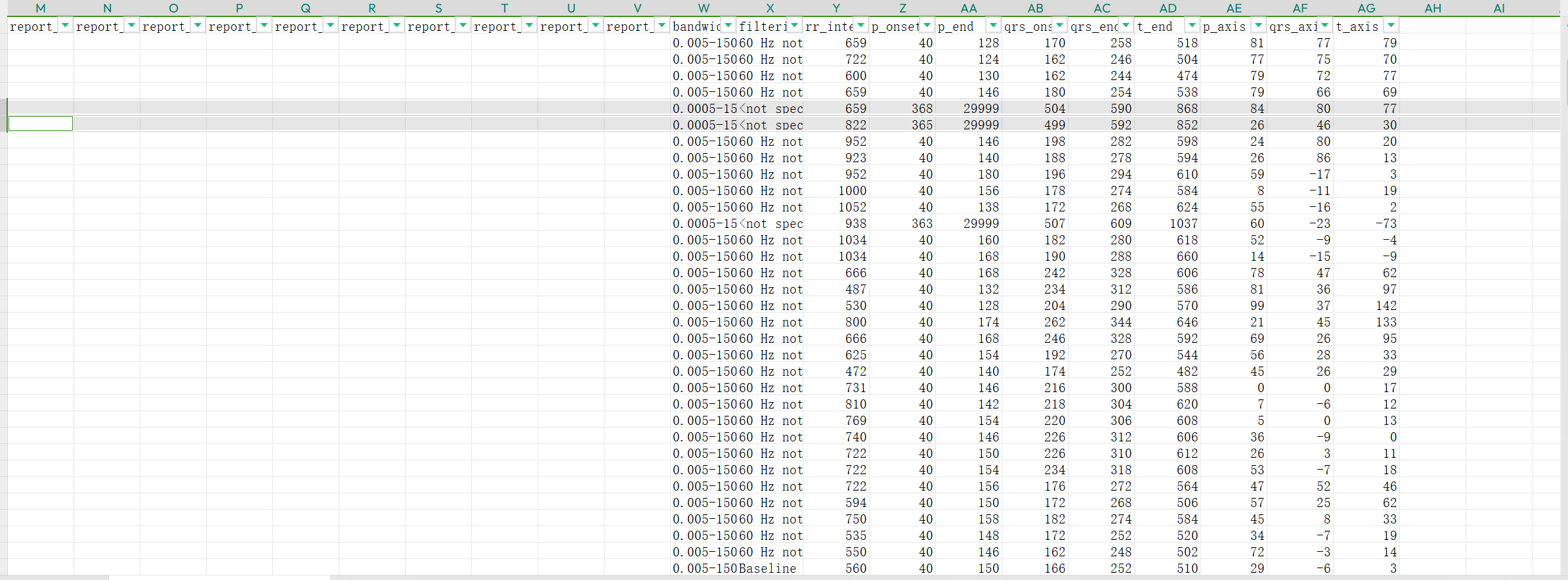


1. machine\_measurements\_data\_dictionary.csv文件中记录了各个机器测量值的描述。



1. machine\_measurements.csv文件中记录了机器测量值以及机器输出的一些report，即machine\_measurements\_data\_dictionary.csv文件中所描述的各类数据的具体值和18次report，同样有16w+患者和80w+数据，但显然部分数据有缺失。具体数据如下：

subject\_id，study\_id，cart\_id，ecg\_time，report\_0，...，report\_17，bandwidth，filtering，rr\_interval，...，t-axis。



可以看出，有缺失值，且偏后位的report多为空（另外，在某些情况下，在带有文本的列之间会有一列没有文本，例如：report\_0：<text\_a>、report\_1：空、report\_2：<text\_b>）

1. waveform\_note\_links.csv文件中记录了subject\_id，study\_id，波形路径，note\_id和记录时间。80w+的数据中，对应了609272条心脏病专家的reports（约20w未能得到报告）。报告通过subject\_id和note\_id对应到Note数据库中的自由文本信息，实现波形和自由文本的链接。其中，note\_id由受试者ID、报告来源域的缩写（EK）和一个顺序整数（顺序整数也列在该文件的 note\_seq 列中，可以用来解读某个受试者在所有就诊过程中收集的心电图的顺序）组成。

Note\_id的格式为：

1. WFDB数据：WFDB（Waveform Database）文件格式通常用于存储生理信号数据，包括头文件（.hea）和数据文件（.dat）两种类型。这两种文件共同构成一个完整的记录：
2. 头文件：一个纯文本文件，提供有关记录的各种信息，如信号的属性、采样率、记录时间等。
3. 数据文件：一个二进制文件，根据头文件中定义的格式和参数，存储信号的原始样本值。该格式可以用于高效存储和处理，但无法直接阅读，但可以通过适当的软件工具进行解析和显示，如python的WFDB工具箱等。（多通道信号的数据通常以交错方式存储，即每个时间点的所有通道数据依次存储，然后是下一个时间点的数据。）

以下为一个样本的全部数据内容，标黄的部分为signal data：

{

'adc\_gain': [200.0, 200.0, 200.0, 200.0, 200.0, 200.0, 200.0, 200.0, 200.0, 200.0, 200.0, 200.0],

'adc\_res': [16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16],

'adc\_zero': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],

'base\_counter': None,

'base\_date': datetime.date(2180, 7, 23),

'base\_time': datetime.time(8, 44),

'baseline': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],

'block\_size': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],

'checksum': [9616, 35318, 10704, 40438, 17944, 7176, 34084, 8828, 30926, 42230, 46797, 31544],

'comments': ['<subject\_id>: 10000032'],

'counter\_freq': None,

'd\_signal': None,

'e\_d\_signal': None,

'e\_p\_signal': None,

'file\_name': ['40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat',

'40689238.dat'],

'fmt': ['16', '16', '16', '16', '16', '16', '16', '16', '16', '16', '16', '16'],

'fs': 500,

'init\_value': [-31, 2, 30, 14, 15, -28, 8, 8, 2, 8, 16, 4],

'n\_sig': 12,

'p\_signal': array([[-0.155, 0.01 , 0.15 , ..., 0.04 , 0.08 , 0.02 ],

[-0.17 , 0.01 , 0.165, ..., 0.03 , 0.08 , 0.02 ],

[-0.17 , 0. , 0.155, ..., 0.025, 0.08 , 0.01 ],

...,

[-0.045, 0.01 , 0.04 , ..., -0.02 , -0.02 , -0.2 ],

[-0.055, 0. , 0.04 , ..., -0.02 , -0.02 , -0.2 ],

[-0.075, -0.01 , 0.05 , ..., -0.02 , -0.02 , -0.2 ]]),

'record\_name': '40689238',

'samps\_per\_frame': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],

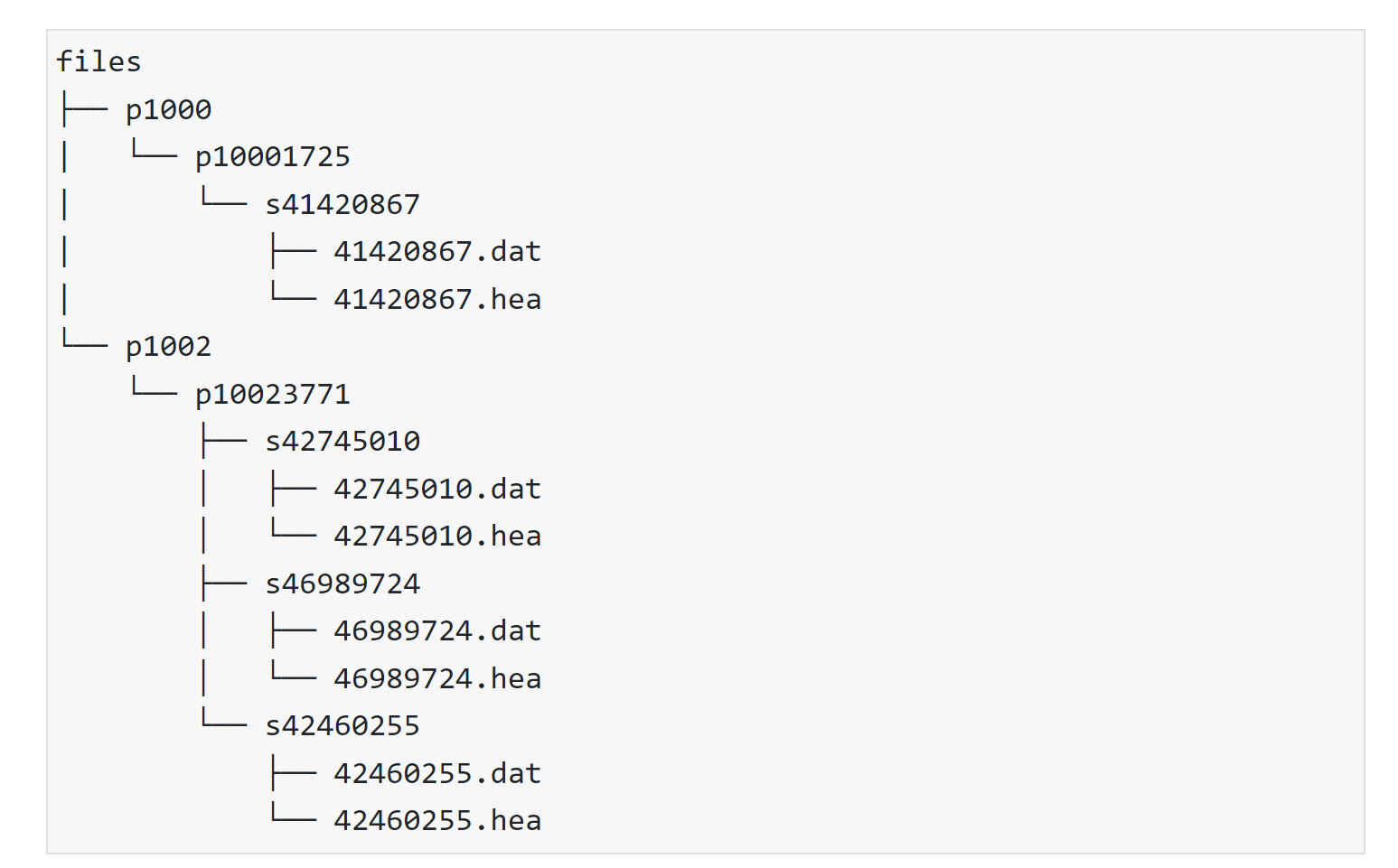
'sig\_len': 5000,

'sig\_name': ['I', 'II', 'III', 'aVR', 'aVF', 'aVL', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6'],

'skew': [None, None, None, None, None, None, None, None, None, None, None, None],

'units': ['mV', 'mV', 'mV', 'mV', 'mV', 'mV', 'mV', 'mV', 'mV', 'mV', 'mV', 'mV']

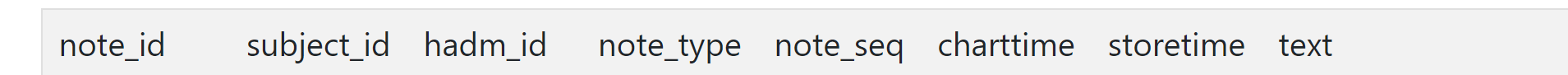
}



# **Note database：**

1、discharge.csv.gz文件

该文件存储了患者的出院摘要（discharge summary），主要包括以下信息：



（1）note\_id: 唯一标识每条note的 ID，通过该ID或subject\_可以与MIMIC IV ECG进行关联。

Note\_id的格式为：

（2）subject\_id: 标识患者的 ID，每个 subject\_id 对应一位患者。

（3）hadm\_id: 住院记录的 ID，标识一次住院事件。每次住院的摘要都与一个特定的 hadm\_id 相关联，类似于ECG数据库中的Study\_ID。

（4）note\_type: 笔记类型，指示该笔记属于哪种类型的医疗记录（例如出院摘要）。

（5）note\_seq: 笔记的顺序号，可能用于标识同一患者在一次住院期间的多条笔记。

（6）charttime: 笔记的创建时间，标识笔记记录的时间点。

（7）storetime: 笔记存储的时间，指示笔记何时保存到数据库中。

（8）text: 笔记的实际内容，包含医生撰写的自由文本形式的出院摘要，描述了患者的入院原因、住院过程以及任何相关的出院指示。Text是该文件的正文内容，格式为一整段文本。

整段文本以类似于字典的形式表示，但键值对之间没有明显的分隔符，常见的键总结如下：

· **患者基本信息**：

* **Name**: 患者姓名（通常已去识别化，一般以下划线表示）
* **Unit No**: 病房号码（已去识别化）
* **Admission Date**: 入院日期
* **Discharge Date**: 出院日期
* **Date of Birth**: 出生日期
* **Sex**: 性别
* **Service**: 医疗服务类型（如 MEDICINE）
* **Allergies**: 过敏信息
* **Attending**: 主治医生（已去识别化）
* **Chief Complaint**: 主诉
* **Major Surgical or Invasive Procedure**: 主要手术或侵入性操作

· **病史与治疗**：

* **History of Present Illness**: 现病史，详细描述患者当前症状和病情发展。
* **Past Medical History**: 既往病史，列举患者之前的疾病。
* **Social History**: 社会史，描述患者的社会背景、生活方式等。
* **Family History**: 家族史，列举患者家庭中的疾病史。
* **Medications on Admission**: 入院时的用药列表。
* **Discharge Medications**: 出院时的用药列表。
* **Discharge Diagnosis**: 出院诊断，列出主要和次要诊断。
* **Discharge Condition**: 出院时的病情状态（如神志、活动状态等）。
* **Discharge Instructions**: 出院指示，提供患者离院后的注意事项和后续护理计划。

· **体格检查与实验室检查**：

* **Physical Exam**: 体格检查结果，包括多个系统的检查，如心血管、呼吸、腹部、神经等。
* **Pertinent Results**: 关键实验室检查结果，通常包括血液检查、成像检查等。
* **Imaging**: 成像检查结果，如CT、X光等。

· **出院总结**：

* **Brief Hospital Course**: 住院期间的简要病情进展。
* **Transitional Issues**: 转诊问题或出院后的随访建议。

第一行示例输出：

Name: \_\_\_ Unit No: \_\_\_

Admission Date: \_\_\_ Discharge Date: \_\_\_

Date of Birth: \_\_\_ Sex: F

Service: MEDICINE

Allergies:

No Known Allergies / Adverse Drug Reactions

Attending: \_\_\_

Chief Complaint:

Worsening ABD distension and pain

Major Surgical or Invasive Procedure:

Paracentesis

History of Present Illness:

\_\_\_ HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD,

bioplar, PTSD, presented from OSH ED with worsening abd

distension over past week.

Pt reports self-discontinuing lasix and spirnolactone \_\_\_ weeks

ago, because she feels like "they don't do anything" and that

she "doesn't want to put more chemicals in her." She does not

follow Na-restricted diets. In the past week, she notes that she

has been having worsening abd distension and discomfort. She

denies \_\_\_ edema, or SOB, or orthopnea. She denies f/c/n/v, d/c,

dysuria. She had food poisoning a week ago from eating stale

cake (n/v 20 min after food ingestion), which resolved the same

day. She denies other recent illness or sick contacts. She notes

that she has been noticing gum bleeding while brushing her teeth

in recent weeks. she denies easy bruising, melena, BRBPR,

hemetesis, hemoptysis, or hematuria.

Because of her abd pain, she went to OSH ED and was transferred

to \_\_\_ for further care. Per ED report, pt has brief period of

confusion - she did not recall the ultrasound or bloodwork at

osh. She denies recent drug use or alcohol use. She denies

feeling confused, but reports that she is forgetful at times.

In the ED, initial vitals were 98.4 70 106/63 16 97%RA

Labs notable for ALT/AST/AP \_\_\_ \_\_\_: \_\_\_,

Tbili1.6, WBC 5K, platelet 77, INR 1.6

Past Medical History:

1. HCV Cirrhosis

2. No history of abnormal Pap smears.

3. She had calcification in her breast, which was removed

previously and per patient not, it was benign.

4. For HIV disease, she is being followed by Dr. \_\_\_ Dr.

\_\_\_.

5. COPD

6. Past history of smoking.

7. She also had a skin lesion, which was biopsied and showed

skin cancer per patient report and is scheduled for a complete

removal of the skin lesion in \_\_\_ of this year.

8. She also had another lesion in her forehead with purple

discoloration. It was biopsied to exclude the possibility of

\_\_\_'s sarcoma, the results is pending.

9. A 15 mm hypoechoic lesion on her ultrasound on \_\_\_

and is being monitored by an MRI.

10. History of dysplasia of anus in \_\_\_.

11. Bipolar affective disorder, currently manic, mild, and PTSD.

12. History of cocaine and heroin use.

Social History:

\_\_\_

Family History:

She a total of five siblings, but she is not

talking to most of them. She only has one brother that she is in

touch with and lives in \_\_\_. She is not aware of any

known GI or liver disease in her family.

Her last alcohol consumption was one drink two months ago. No

regular alcohol consumption. Last drug use \_\_\_ years ago. She

quit smoking a couple of years ago.

Physical Exam:

VS: 98.1 107/61 78 18 97RA

General: in NAD

HEENT: CTAB, anicteric sclera, OP clear

Neck: supple, no LAD

CV: RRR,S1S2, no m/r/g

Lungs: CTAb, prolonged expiratory phase, no w/r/r

Abdomen: distended, mild diffuse tenderness, +flank dullness,

cannot percuss liver/spleen edge \_\_\_ distension

GU: no foley

Ext: wwp, no c/e/e, + clubbing

Neuro: AAO3, converse normally, able to recall 3 times after 5

minutes, CN II-XII intact

Discharge:

PHYSICAL EXAMINATION:

VS: 98 105/70 95

General: in NAD

HEENT: anicteric sclera, OP clear

Neck: supple, no LAD

CV: RRR,S1S2, no m/r/g

Lungs: CTAb, prolonged expiratory phase, no w/r/r

Abdomen: distended but improved, TTP in RUQ,

GU: no foley

Ext: wwp, no c/e/e, + clubbing

Neuro: AAO3, CN II-XII intact

Pertinent Results:

\_\_\_ 10:25PM GLUCOSE-109\* UREA N-25\* CREAT-0.3\* SODIUM-138

POTASSIUM-3.4 CHLORIDE-105 TOTAL CO2-27 ANION GAP-9

\_\_\_ 10:25PM estGFR-Using this

\_\_\_ 10:25PM ALT(SGPT)-100\* AST(SGOT)-114\* ALK PHOS-114\*

TOT BILI-1.6\*

\_\_\_ 10:25PM LIPASE-77\*

\_\_\_ 10:25PM ALBUMIN-3.3\*

\_\_\_ 10:25PM WBC-5.0# RBC-4.29 HGB-14.3 HCT-42.6 MCV-99\*

MCH-33.3\* MCHC-33.5 RDW-15.7\*

\_\_\_ 10:25PM NEUTS-70.3\* LYMPHS-16.5\* MONOS-8.1 EOS-4.2\*

BASOS-0.8

\_\_\_ 10:25PM PLT COUNT-71\*

\_\_\_ 10:25PM \_\_\_ PTT-30.9 \_\_\_

\_\_\_ 10:25PM \_\_\_

.

CXR: No acute cardiopulmonary process.

U/S:

1. Nodular appearance of the liver compatible with cirrhosis.

Signs of portal

hypertension including small amount of ascites and splenomegaly.

2. Cholelithiasis.

3. Patent portal veins with normal hepatopetal flow.

Diagnostic para attempted in the ED, unsuccessful.

On the floor, pt c/o abd distension and discomfort.

Brief Hospital Course:

\_\_\_ HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD,

bioplar, PTSD, presented from OSH ED with worsening abd

distension over past week and confusion.

# Ascites - p/w worsening abd distension and discomfort for last

week. likely \_\_\_ portal HTN given underlying liver disease,

though no ascitic fluid available on night of admission. No

signs of heart failure noted on exam. This was \_\_\_ to med

non-compliance and lack of diet restriction. SBP negative

diuretics:

> Furosemide 40 mg PO DAILY

> Spironolactone 50 mg PO DAILY, chosen over the usual 100mg

dose d/t K+ of 4.5.

CXR was wnl, UA negative, Urine culture blood culture negative.

Pt was losing excess fluid appropriately with stable lytes on

the above regimen. Pt was scheduled with current PCP for

\_\_\_ check upon discharge.

Pt was scheduled for new PCP with Dr. \_\_\_ at \_\_\_ and

follow up in Liver clinic to schedule outpatient screening EGD

and \_\_\_.

Medications on Admission:

The Preadmission Medication list is accurate and complete.

1. Furosemide 20 mg PO DAILY

2. Spironolactone 50 mg PO DAILY

3. Albuterol Inhaler 2 PUFF IH Q4H:PRN wheezing, SOB

4. Raltegravir 400 mg PO BID

5. Emtricitabine-Tenofovir (Truvada) 1 TAB PO DAILY

6. Nicotine Patch 14 mg TD DAILY

7. Ipratropium Bromide Neb 1 NEB IH Q6H SOB

Discharge Medications:

1. Albuterol Inhaler 2 PUFF IH Q4H:PRN wheezing, SOB

2. Emtricitabine-Tenofovir (Truvada) 1 TAB PO DAILY

3. Furosemide 40 mg PO DAILY

RX \*furosemide 40 mg 1 tablet(s) by mouth Daily Disp #\*30 Tablet

Refills:\*3

4. Ipratropium Bromide Neb 1 NEB IH Q6H SOB

5. Nicotine Patch 14 mg TD DAILY

6. Raltegravir 400 mg PO BID

7. Spironolactone 50 mg PO DAILY

8. Acetaminophen 500 mg PO Q6H:PRN pain

Discharge Disposition:

Home

Discharge Diagnosis:

Ascites from Portal HTN

Discharge Condition:

Mental Status: Clear and coherent.

Level of Consciousness: Alert and interactive.

Activity Status: Ambulatory - Independent.

Discharge Instructions:

Dear Ms. \_\_\_,

It was a pleasure taking care of you! You came to us with

stomach pain and worsening distension. While you were here we

did a paracentesis to remove 1.5L of fluid from your belly. We

also placed you on you 40 mg of Lasix and 50 mg of Aldactone to

help you urinate the excess fluid still in your belly. As we

discussed, everyone has a different dose of lasix required to

make them urinate and it's likely that you weren't taking a high

enough dose. Please take these medications daily to keep excess

fluid off and eat a low salt diet. You will follow up with Dr.

\_\_\_ in liver clinic and from there have your colonoscopy

and EGD scheduled. Of course, we are always here if you need us.

We wish you all the best!

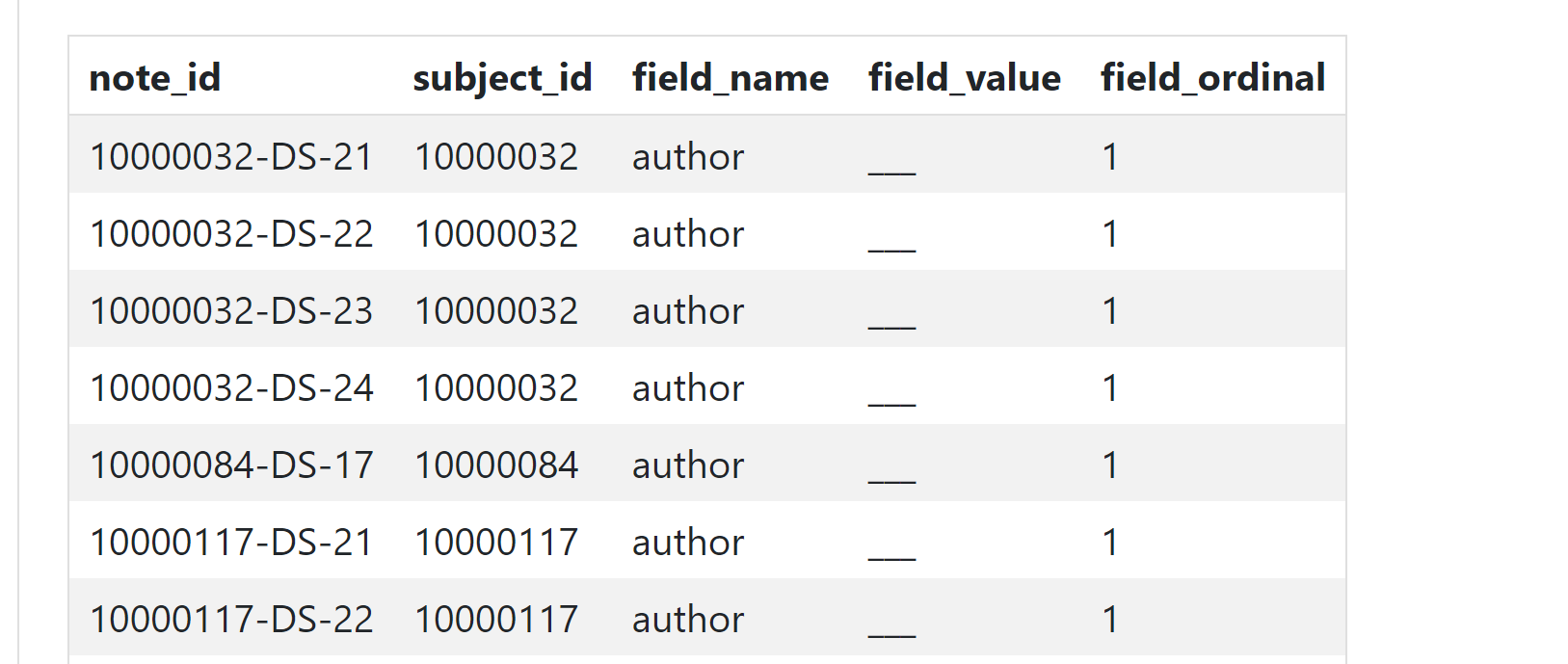
Your \_\_\_ Team.

Followup Instructions:

\_\_\_

2、discharge\_detail.csv.gz文件：

包含与出院摘要相关的辅助信息，包括出院摘要作者的去识别化占位符，体现对隐私的保护，对研究没有作用：



field\_name: 该字段表示记录的属性名称。在本例中，author 表示出院摘要的作者信息。

field\_value: 这是与field\_name相关的实际数据值。在这里，它存储的是作者信息的去识别化占位符，即“\_\_\_”。

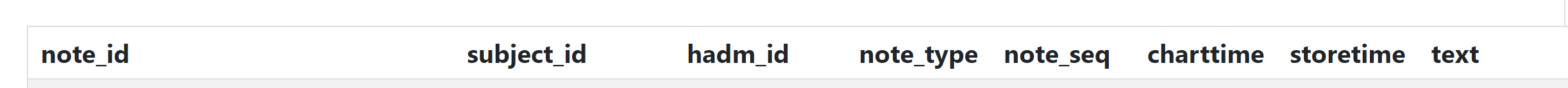
dinal: 该字段表示同一field\_name的顺序号。在这里，所有的记录都为1，表示这是唯一的值。

此表格的用途是记录与出院摘要相关的附加信息，并通过note\_id将这些信息与具体的出院摘要文本关联起来。

3、radiology.csv.gz文件：

该文件包含与放射成像相关的自由文本放射报告。放射报告涵盖多种成像模式：X射线、计算机断层扫描（CT）、磁共振成像（MRI）、超声等。自由文本放射报告是半结构化的，通常根据特定的成像协议遵循一致的模板。例如，胸部X光报告通常包括四个部分：适应症、对比、发现和印象。

其表头与discharge summary一样，原理也一样，只是note\_type由DS（discharge summary）变成了RR（radiology report），note列的自由文本内容相应发生变化。



text列为一整段文本，但依然存在类似字典的形式，同文件1，常见的键如下：

· EXAMINATION：检查类型或部位

· CLINICAL INFORMATION / INDICATION：临床信息或检查的指征

· COMPARISON：与之前检查的对比

· TECHNIQUE：使用的技术或方法

· FINDINGS：检查的发现

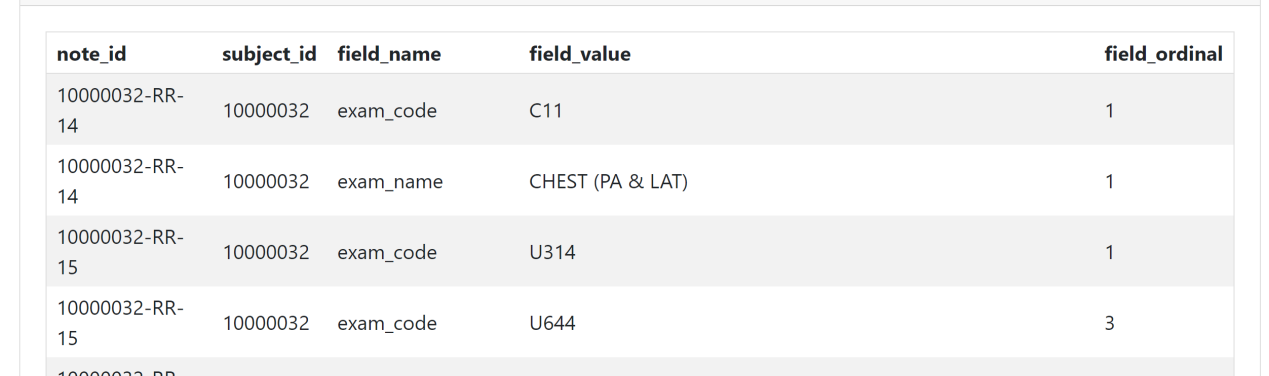
· PROCEDURE：具体操作和步骤（如果有）

· SUPINE AND UPRIGHT VIEWS OF THE ABDOMEN：腹部仰卧和直立视图（特定于腹部影像）

· IMPRESSION：医生的结论或印象

4、radiology.csv\_details.gz文件：

该文件提供与影像研究相关的信息。该表中包括当前程序术语（CPT）代码、检查名称以及主报告与附加报告之间的链接。



· note\_id: 这是唯一标识每一条记录的ID，通常由患者ID、笔记领域（例如“RR”代表放射学报告）以及一个顺序整数组成，用于关联不同表之间的数据。

· subject\_id: 这是患者的唯一标识符，用于识别与特定患者相关的所有数据。例如，这里10000032 代表特定的患者。

· field\_name: 该字段表示记录的属性名称。在这个文件中，可能的field\_name包括 exam\_code 和 exam\_name。exam\_code代表检查的CPT代码，而 exam\_name 代表检查的名称。

· field\_value: 这是与 field\_name 相关的实际数据值。这里，field\_value 包含了检查的代码（如 C11）和检查的名称（如 CHEST (PA & LAT)）。

· field\_ordinal: 该字段表示同一 field\_name 的顺序号。这对于某些情况下的多值字段很有用，例如当一个检查包含多个程序代码时。



# Emergency department database：

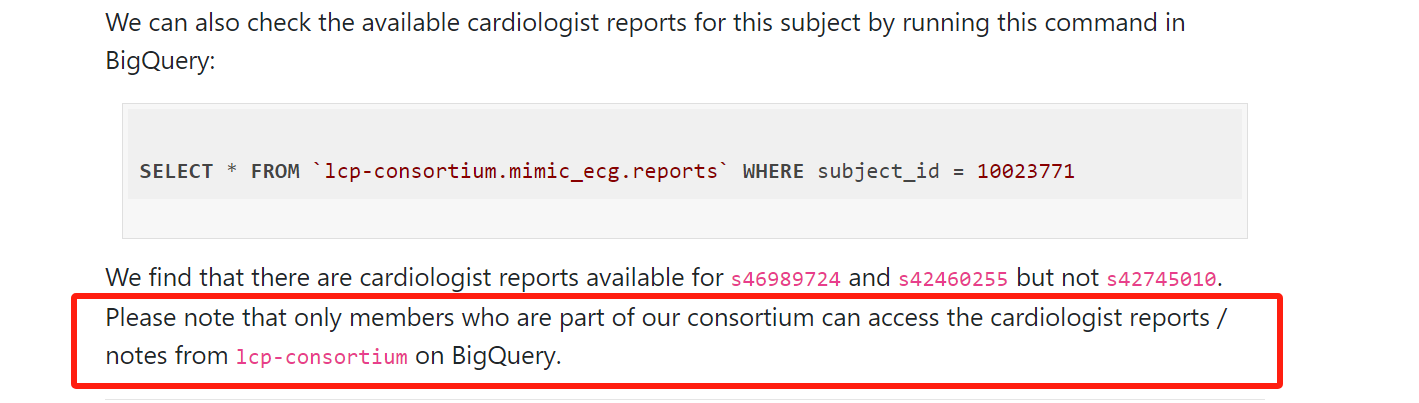
8.20 ECG和NOTE数据库只能对齐心电图、心电图患者、患者的出院摘要。

然而，出院摘要是一整段大文本，什么病都有，大多没有心脏相关的描述，即使有，也需要手动检查，而且不确定“出院摘要里有关心脏病的描述”是否和“心脏病专家报告的描述”是一致的，以及是否有疾病的标签（或称为样本）。下一步的两个方向：

1、复现论文，看看别人是怎么提取数据和标签、对齐多模态数据的

2、重新检查icu/hosp 数据库、ED数据库，寻找ECG相关或心脏病专家报告相关的数据

3、使用BigQuery，注册账号检查龙慧姐的physionet账号是否加入了协会



指引：

<https://mimic.mit.edu/docs/gettingstarted/cloud/bigquery/>