2026년 대한부인종양학회
제7회 동계학술대회 with Chemo-TIP Review
일자 2026년 1월 17일 (토)
장소 세종대학교 컨벤션센터

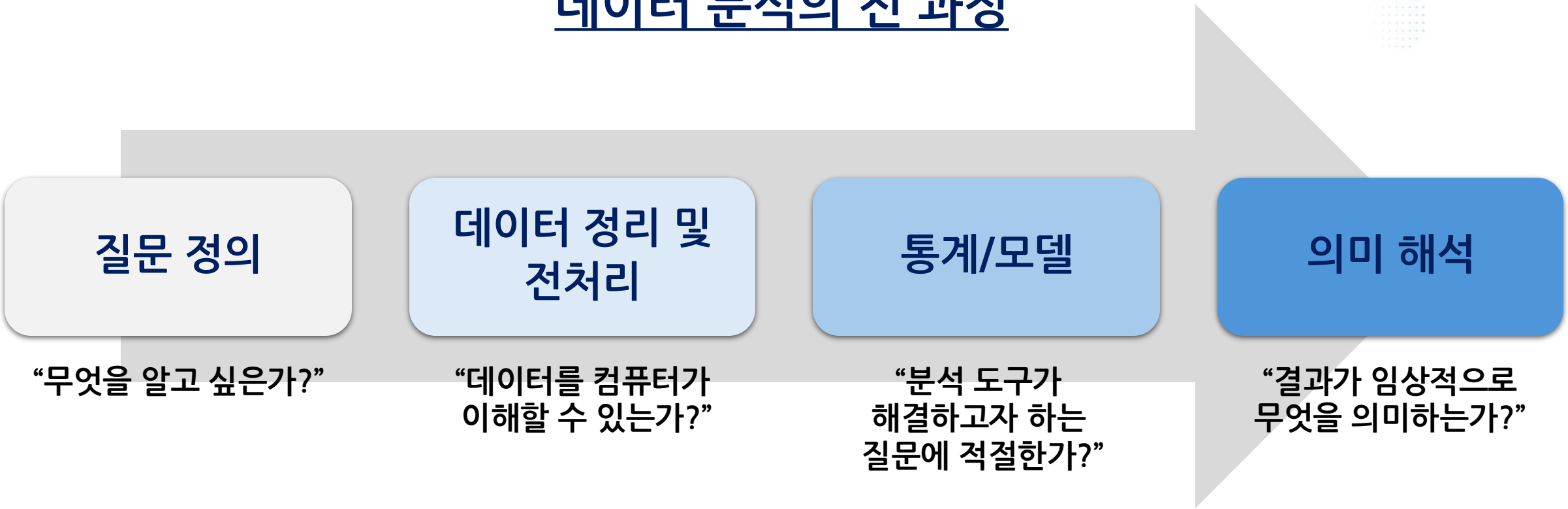
Comparison of Data Analysis Tools: Excel, R, and Python

Yujeong Kim, Ph.D
(Department of Biomedical Systems Informatics,
College of medicine, Yonsei University)

Table of Contents

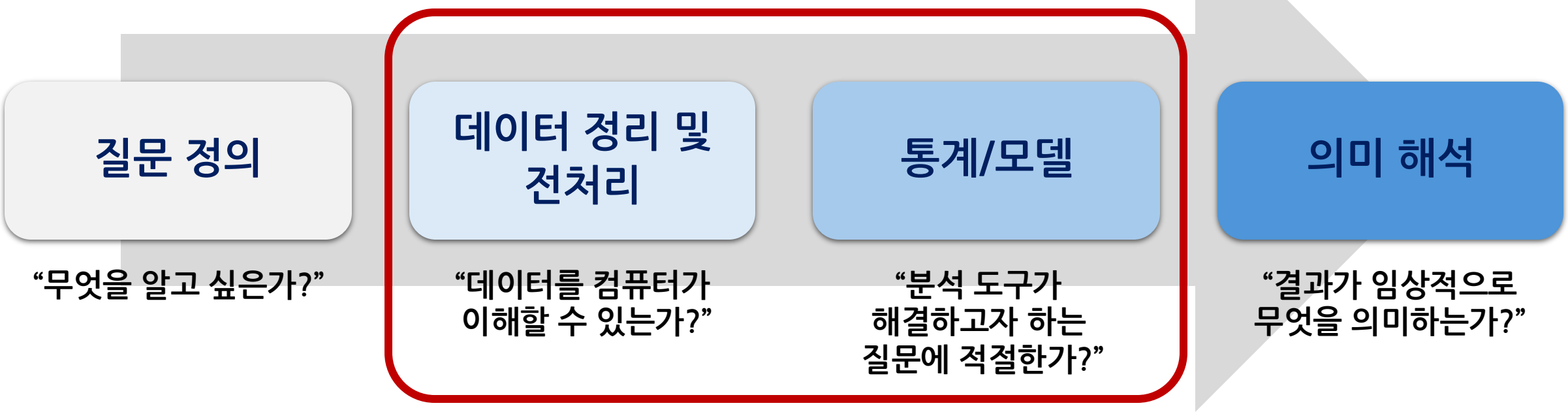
‘데이터 분석’ 이란?

데이터 분석의 전 과정



'데이터 분석' 이란?

데이터 분석의 전 과정



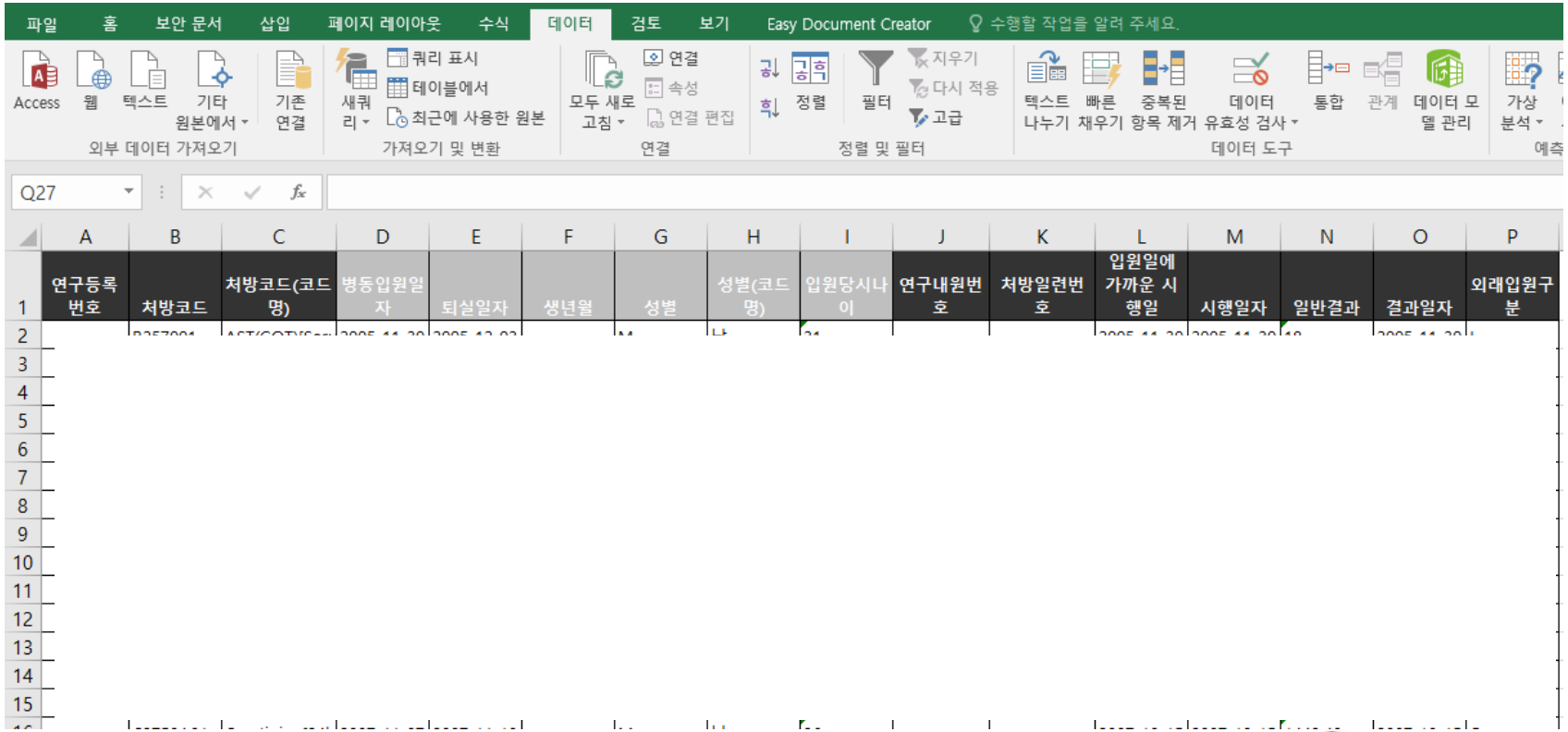
데이터 분석 tool

- Excel
  - 표를 가장 쉽게 다룰 수 있는 tool (exploratory tool)
  - 데이터를 처음 접하고, 구조를 이해하는 단계에 적합
- R
  - 의학 통계의 기본. 통계 분석의 강자
  - 임상 연구에서 가장 표준적인 통계 언어
- Python
  - AI/예측 모델을 위한 powerful한 도구
  - 다양한 형태의 임상 데이터를 하나로 연결하는 도구

Excel 기반의 데이터 분석

Excel: 표를 가장 쉽게 다룰 수 있는 tool (exploratory tool)

- 분석을 '직접' 수행하는 도구라기 보단, 분석을 '가능'하게 만드는 입력/탐색적 도구
  - 데이터 구조 확인
  - 오류, 결측치 발견
  - 간단한 요약과 분포 확인
  - 데이터 정렬, 필터링



R 기반의 데이터 분석

R: 의학 통계의 기본. 통계의 강자

- 질문에 대한 통계적 답을 제시하는 도구
  - Excel에서 가능한 데이터 정리, 요약 기능들
  - 가설 검정, 회귀 분석, 생존 분석, 군 매칭 등 강력한 통계 분석
  - 최근에는 머신러닝 까지도 가능
  - 논문용 figure, table 제작

  - Excel은 보통 한 파일(시트) 중심으로 작업하지만,
    R은 여러 파일을 한 번에 불러와 같은 프로젝트 안에서
    원자료는 유지한 채 “분석용 데이터셋(부분집합)”을 여러 개 만들어 저장하고 비교할 수 있음

R 기반의 데이터 분석

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Untitled1*

Source on Save | Run | Source

```
1  data <-read.csv("Z:/research/data_for_ML_fn_v1_co.csv", fileEncoding="CP949", check.
2
3  age_over60 <- subset(data, 입원당시나이>=60)
4  age_less60 <- subset(data, 입원당시나이<60)
5
6
```

Environment | History | Connections | Tutorial

Import Dataset | 132 MiB | List

R | Global Environment

Data

age_less60	1439 obs. of 12 variables
age_over60	156 obs. of 12 variables
data	1595 obs. of 12 variables

Untitled1* | data | age_over60 | age_less60

Filter

	연구등록번호	병리번호	병동입원일자	성별	입원당시나이	측정값_DBP	항목값_Drink	측정값_Height	측정값_SBP	항목값_Smoke	측
1			2021-01-10	M	38	88	nondrinker	178.1	144	nonsmoker	
2			2018-12-04	F	30	85	current drinker	163.6	119	nonsmoker	
3			2020-12-03	F	47	79	nondrinker	160.5	121	nonsmoker	
4			2020-04-02	M	36	75	current drinker	177.4	114	nonsmoker	
5			2007-01-12	M	24	70	유	178.0	110	nonsmoker	
6			2019-11-12	F	50	78	ex-drinker	168.0	118	nonsmoker	
7			2019-02-21	F	49	65	nondrinker	183.8	100	nonsmoker	
8			2020-04-09	M	27	65	nondrinker	176.9	114	ex-smoker	
9			2018-11-22	M	50	84	ex-drinker	166.0	135	nonsmoker	
10			2017-08-23	F	68	90	nondrinker	149.9	156	nonsmoker	

R 기반의 데이터 분석

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Untitled1* | age_over60 | data | imputation.R | OA_T2DM.R

Source on Save | Run | Source

```
1  data <-read.csv("Z:/research/data_for_ML_fn_v1_co.csv", fileEncoding="CP949", check.na
2
3  age_over60 <- subset(data, 입원당시나이>=60)
4  age_less60 <- subset(data, 입원당시나이<60)
5
6  par(mfrow = c(1, 2))
7
8  hist(age_over60$측정값_DBP, main = "Age ≥ 60",
9       xlab = "DBP",
10      col = "lightblue")
11
12 hist(age_less60$측정값_DBP,
13      main = "Age < 60",
14      xlab = "DBP",
15      col = "lightgreen")
16
```

16:1 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.4.0 · ~/

```
> hist(age_over60$측정값_DBP, main = "Age ≥ 60",
+      xlab = "DBP",
+      col = "lightblue")
>
> hist(age_less60$측정값_DBP,
+      main = "Age < 60",
+      xlab = "DBP",
+      col = "lightgreen")
>
```

Environment | History | Connections | Tutorial

Import Dataset | 138 MiB | List

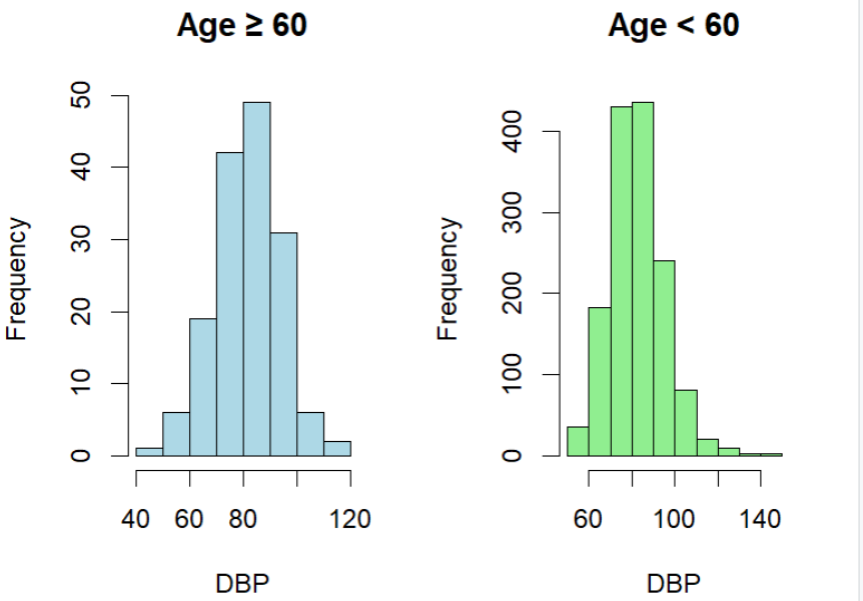
R | Global Environment

Data

age_less60	1439 obs. of 12 variables
age_over60	156 obs. of 12 variables
data	1595 obs. of 12 variables

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export | Publish



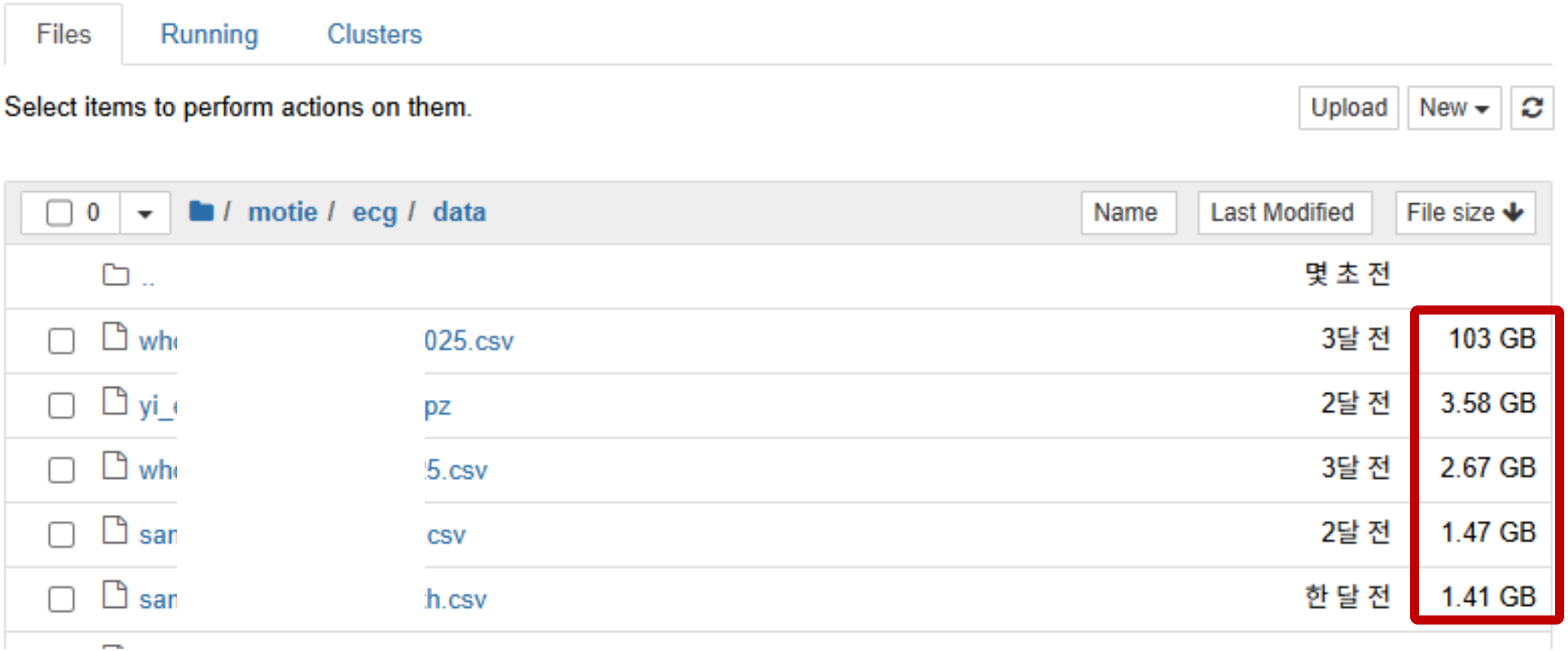
Python 기반의 데이터 분석

Python: AI/예측 모델을 위한 powerful한 tool

- 데이터를 분석→모델·시스템으로 확장하는 도구
  - Excel/R에서 정리된 데이터를 대규모로 처리·자동화
  - 표 데이터 + 이미지 + 신호 + 텍스트 등 다양한 형태의 데이터 분석 가능
  - 머신러닝·딥러닝 기반 예측 모델 개발에 강점
  - 분석 과정을 코드로 고정하여 반복 실행·재현·확장 가능

  - Excel/R은 주로 분석 중심이라면,
    Python은 **분석 + 전처리 + 자동화 + 모델링**을 하나의 파이프라인으로 구성 가능

Python 기반의 데이터 분석



대용량의 데이터도 무리 없이 활용 가능

Python 기반의 데이터 분석

```
ecg=pd.read_csv('
plt.plot(ecg['II'])
```

```
[<matplotlib.lines.Line2D at 0x7f00f2caf130>]
```

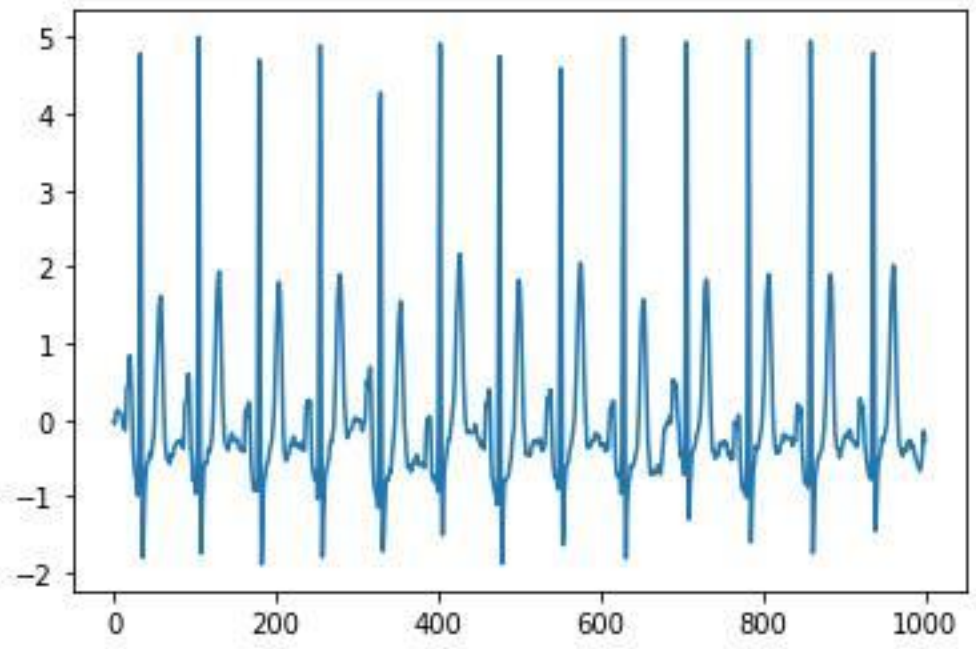


표 데이터뿐 아니라 ECG와 같은 생체 신호를
직접 불러와 시각화·분석 가능

Python 기반의 데이터 분석

```python
filenames = df['filter100'].tolist()

data_list=[]
for filename in filenames :
    data = pd.read_csv(f'/home/Dementia_final/raw_NN/data/filtered100_ecg/{filename}')
    data = data[['I', 'II','III','aVR', 'aVL', 'aVF', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6']]
    data_list.append(data.values)
```

```python
len(data_list)
```

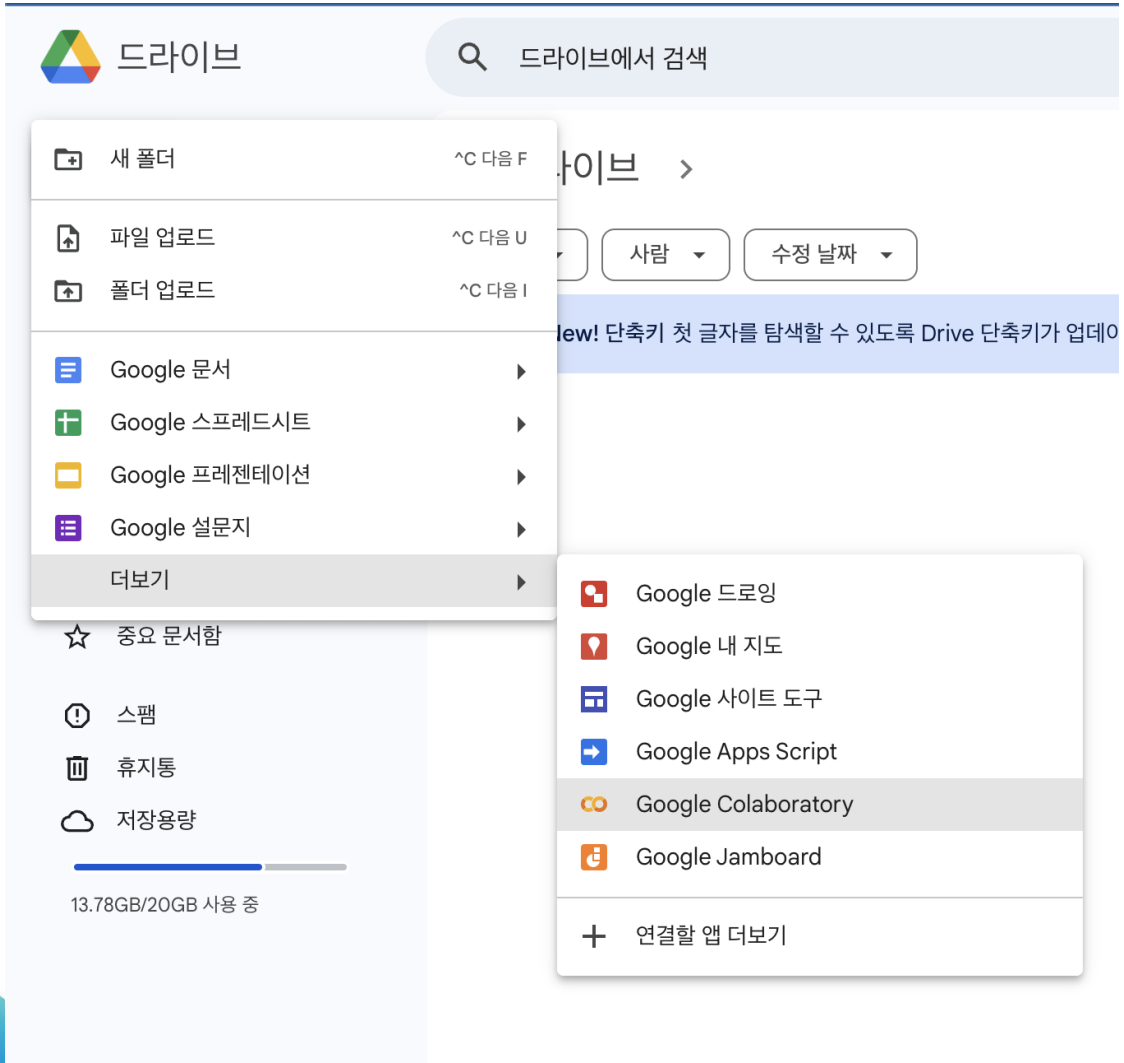
```
10186
```

수 천, 수 만 건 이상의 반복 연산을
효율적으로 수행 가능

```python
from tqdm import tqdm
import pandas as pd

ECG_FOLDER = '/home/motie/ecg/data/filtered_sampled_re_100hz/'  # ECG CSV들이 들어 있는 폴더
# ECG_FOLDER = '/home/motie/ecg/data/filtered_sampled_re/'  # ECG CSV들이 들어 있는 폴더

all_results = []

# df['file'] 컬럼의 각 파일명에 대해 반복
for file_name in tqdm(total_clear['file'], desc="🔍 ECG 파일 추론 중", unit="file"):
    csv_path = os.path.join(ECG_FOLDER, file_name)

    # 파일 존재 여부 확인 (안전장치)
    if not os.path.exists(csv_path):
        print(f"⚠️ 파일 없음: {csv_path}")
        continue

    # 예측 수행
    probs = predict_from_csv(csv_path)

    # 결과 저장
    df_temp = pd.DataFrame([probs], columns=label_texts)
    df_temp.insert(0, "file", file_name)
    all_results.append(df_temp)

# 모든 결과 합치기
df_result = pd.concat(all_results, ignore_index=True)
```

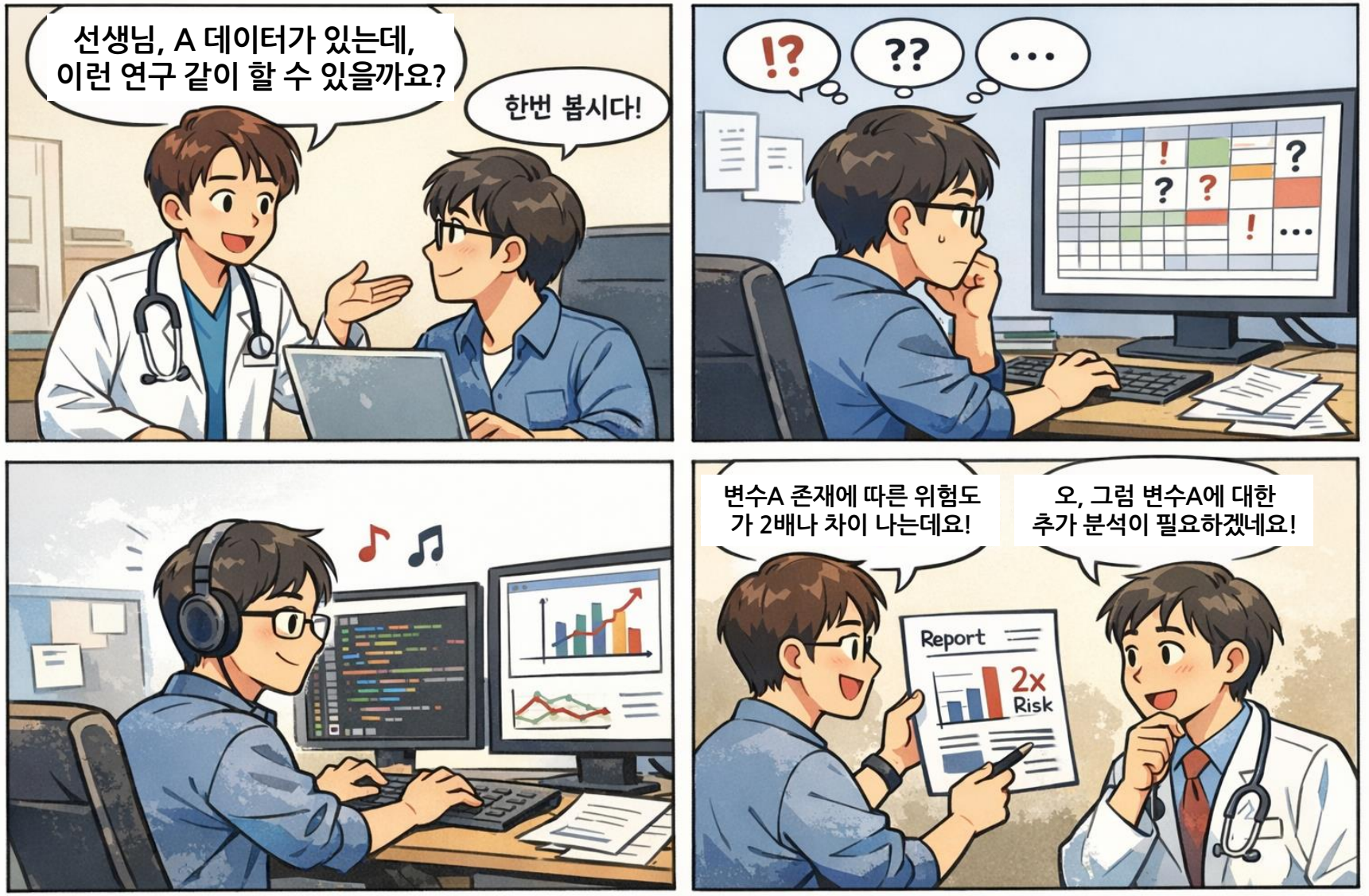
```
🔍 ECG 파일 추론 중: 100%|██████████| 6110/6110 [02:10<00:00, 46.66file/s]
```

Python 기반의 데이터 분석

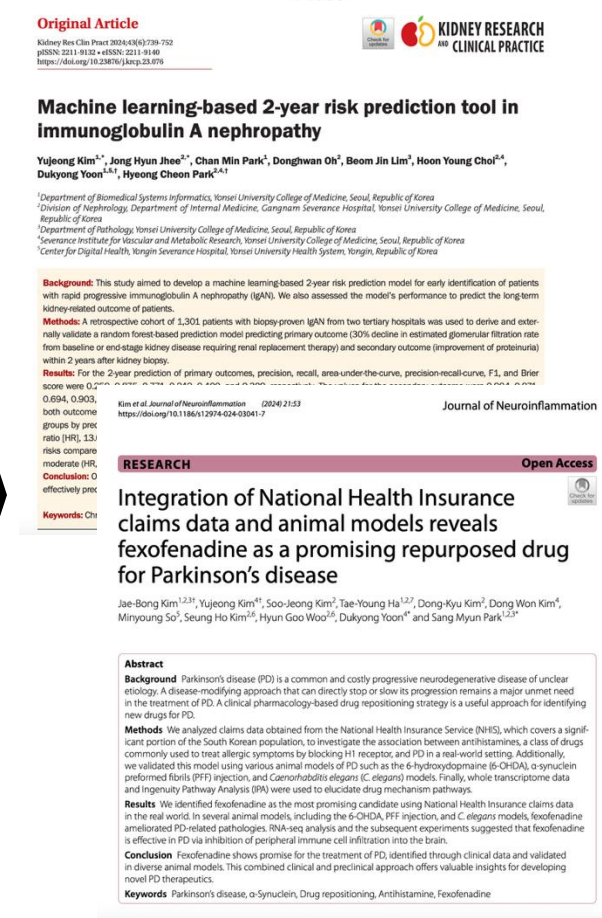


- 설치 없이 쉽게 Python을 실행할 수 있는 방법
- Colaboratory (Colab): Google에서 제공, 설치 없이 Python 코드를 실행할 수 있는 환경
  - 본인의 Google drive와 연결하여 사용이 용이함 (단, 환자 개인 식별 정보는 업로드하지 않고, 연구·교육·예제용 비식별화 데이터 위주로 사용)

데이터사이언티스트와의 데이터 분석 과정



Generated by ChatGPT



Original Article

Kidney Res Clin Pract 2024;43(6):739-752
pISSN: 2211-9132 • eISSN: 2211-9140
https://doi.org/10.23876/j.krcp.23.076

KIDNEY RESEARCH AND CLINICAL PRACTICE

Machine learning-based 2-year risk prediction tool in immunoglobulin A nephropathy

Yujeong Kim[1,*], Jong Hyun Jhee[2,*], Chan Min Park[1], Donghwan Oh[2], Beom Jin Lim[3], Hoon Young Choi[2,4], Dukyong Yoon[1,5,†], Hyeong Cheon Park[2,4,†]

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea
[2]Division of Nephrology, Department of Internal Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea
[3]Department of Pathology, Yonsei University College of Medicine, Seoul, Republic of Korea
[4]Severance Institute for Vascular and Metabolic Research, Yonsei University College of Medicine, Seoul, Republic of Korea
[5]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Republic of Korea

Background: This study aimed to develop a machine learning-based 2-year risk prediction model for early identification of patients with rapid progressive immunoglobulin A nephropathy (IgAN). We also assessed the model's performance to predict the long-term kidney-related outcome of patients.

Methods: A retrospective cohort of 1,301 patients with biopsy-proven IgAN from two tertiary hospitals was used to derive and externally validate a random forest-based prediction model predicting primary outcome (30% decline in estimated glomerular filtration rate from baseline or end-stage kidney disease requiring renal replacement therapy) and secondary outcome (improvement of proteinuria) within 2 years after kidney biopsy.

Results: For the 2-year prediction of primary outcomes, precision, recall, area-under-the-curve, precision-recall-curve, F1, and Brier score were 0.

0.694, 0.903,

both outcome

groups by pre

ratio [HR], 13

risks compare

moderate (HR,

Conclusion: O

effectively pre

Keywords: Ch

Kim et al. Journal of Neuroinflammation (2024) 21:53
https://doi.org/10.1186/s12974-024-03041-7

Journal of Neuroinflammation

RESEARCH Open Access

Integration of National Health Insurance claims data and animal models reveals fexofenadine as a promising repurposed drug for Parkinson's disease

Jae-Bong Kim[1,2,3†], Yujeong Kim[4†], Soo-Jeong Kim[2], Tae-Young Ha[1,2,7], Dong-Kyu Kim[2], Dong Won Kim[4], Minyoung So[5], Seung Ho Kim[2,6], Hyun Goo Woo[2,6], Dukyong Yoon[4*] and Sang Myun Park[1,2,3*]

Abstract

Background Parkinson's disease (PD) is a common and costly progressive neurodegenerative disease of unclear etiology. A disease-modifying approach that can directly stop or slow its progression remains a major unmet need in the treatment of PD. A clinical pharmacology-based drug repositioning strategy is a useful approach for identifying new drugs for PD.

Methods We analyzed claims data obtained from the National Health Insurance Service (NHIS), which covers a significant portion of the South Korean population, to investigate the association between antihistamines, a class of drugs commonly used to treat allergic symptoms by blocking H1 receptor, and PD in a real-world setting. Additionally, we validated this model using various animal models of PD such as the 6-hydroxydopamine (6-OHDA), α-synuclein preformed fibrils (PFF) injection, and *Caenorhabditis elegans* (*C. elegans*) models. Finally, whole transcriptome data and Ingenuity Pathway Analysis (IPA) were used to elucidate drug mechanism pathways.

Results We identified fexofenadine as the most promising candidate using National Health Insurance claims data in the real world. In several animal models, including the 6-OHDA, PFF injection, and *C. elegans* models, fexofenadine ameliorated PD-related pathologies. RNA-seq analysis and the subsequent experiments suggested that fexofenadine is effective in PD via inhibition of peripheral immune cell infiltration into the brain.

Conclusion Fexofenadine shows promise for the treatment of PD, identified through clinical data and validated in diverse animal models. This combined clinical and preclinical approach offers valuable insights for developing novel PD therapeutics.

Keywords Parkinson's disease, α-Synuclein, Drug repositioning, Antihistamine, Fexofenadine

데이터사이언티스트와의 협업

가장 좋은 데이터 = Tidy data!

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

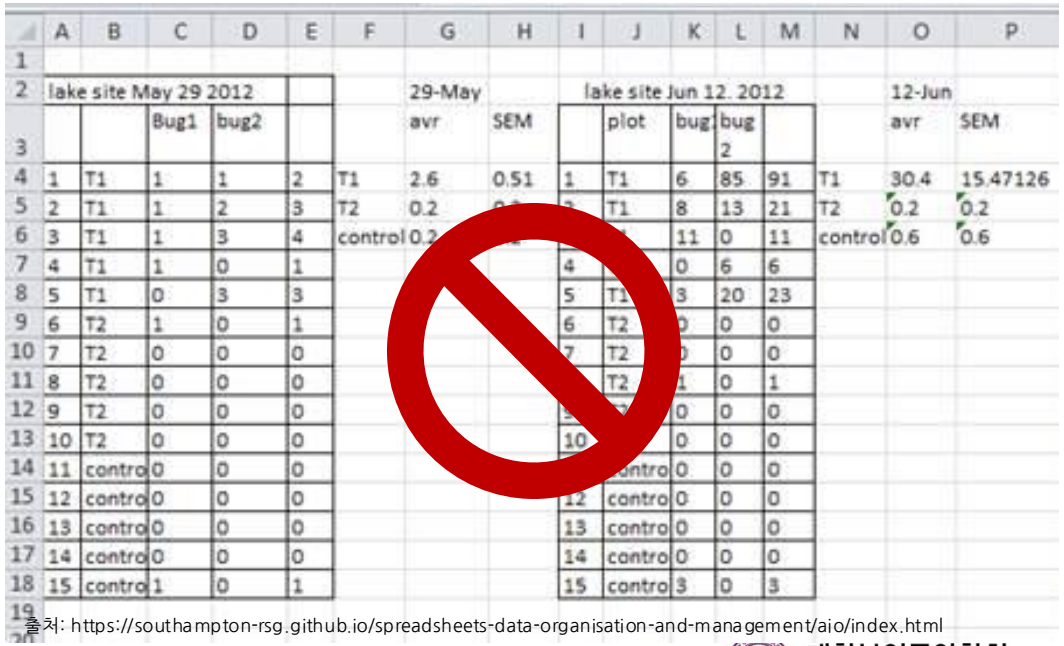
each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

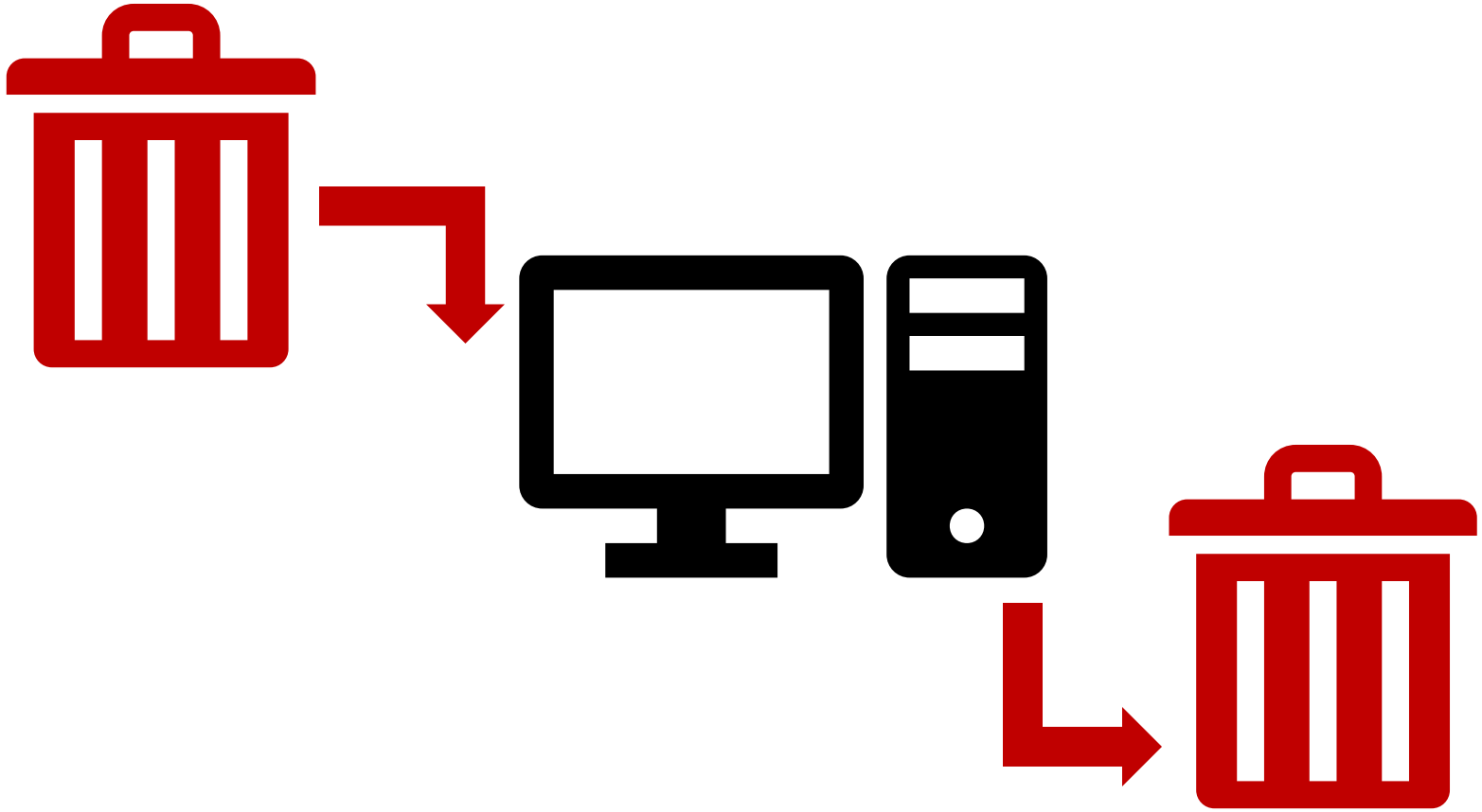
- 변수 1개 = 열 1개, 관측치 1개 = 행 1개
- 한 셀에는 한 정보만! (Male/HTN X)
- 셀 병합은 없도록



출처: https://southampton-rsg.github.io/spreadsheets-data-organisation-and-management/aio/index.html

데이터사이언티스트와의 협업

- 데이터와 함께 제공되면 좋은 자료들
  - 자료 정의서 (Codebook)
    - ✓ 각 변수들이 어떤 것을 의미하는지 (구분 key가 무엇인지, patient_id? Or visit_id?)
    - ✓ 결측치의 존재 유무
    - ✓ 이상치 해결법
    - ✓ 데이터 수집 기간
    - ✓ 반복 측정 여부
  - 분석 요청서
    - ✓ 연구의 main question
    - ✓ Primary (or secondary) outcome 정의
    - ✓ Inclusion/exclusion criteria



Garbage In Garbage Out

감 사 합 니 다

Yujeong Kim, Ph.D
yjkim9346@yuhs.ac

www.sgo.or.kr

2026년 대한부인종양학회 제7회 동계학술대회 with Chemo-TIP Review

일자 2026년 1월 17일 (토)

장소 세종대학교 컨벤션센터

Thank you for your attention!

대한부인종양학회
Korean Society of Gynecologic Oncology