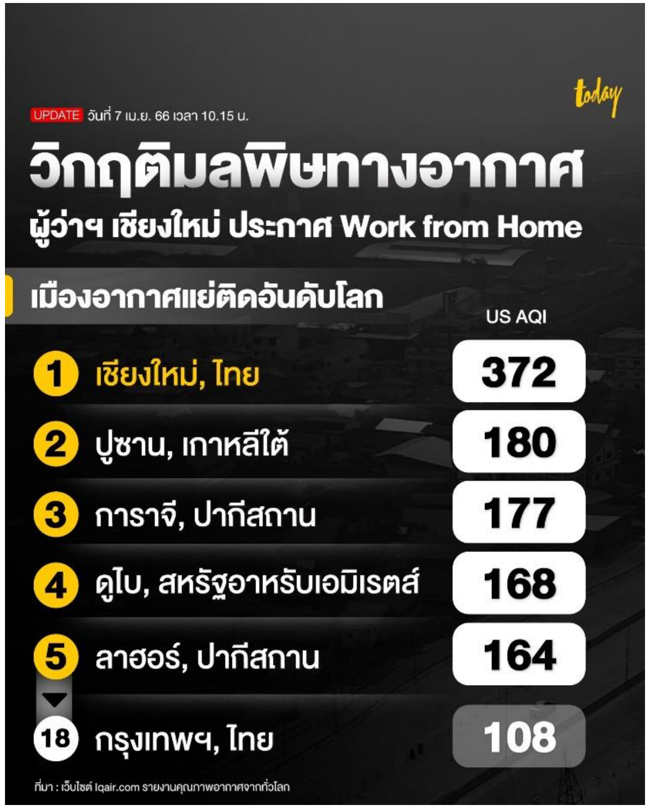
Spatial air pollution and health effects
มลภาวะทางอากาศเชิงพื้นที่และผลต่อสุขภาพ

นพ.ปวิน นำธวัช ภาควิชาระบาดวิทยาคลินิกและชีวสถิติ คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี
ราชวิทยาลัยอายุรแพทย์แห่งประเทศไทย ร่วมกับ สปสช.

Introduction

Ambient air pollution accounts for approximately 4.2 million deaths globally.

worsening in low- and middle-income countries (LMIC) in many regions of the world: Eastern Mediterranean, South-East Asia, and Western Pacific.



today

UPDATE วันที่ 7 เม.ย. 66 เวลา 10.15 น.

วิกฤติมลพิษทางอากาศ

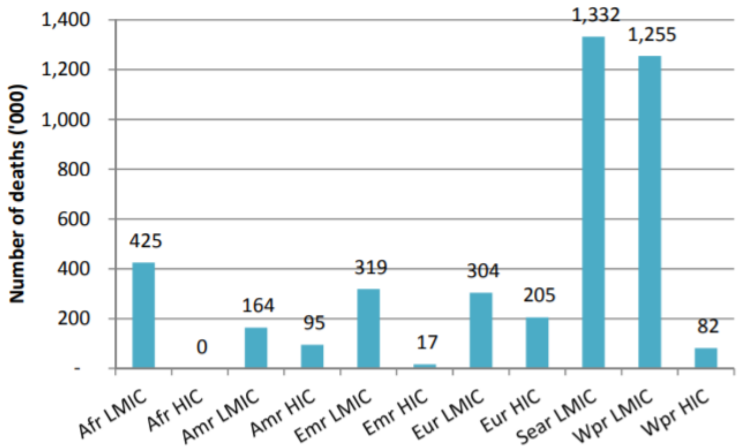
ผู้ว่าฯ เชียงใหม่ ประกาศ Work from Home

เมืองอากาศแย่ติดอันดับโลก

		US AQI
1	เชียงใหม่, ไทย	372
2	ปูซาน, เกาหลีใต้	180
3	การาจี, ปากีสถาน	177
4	ดูไบ, สหรัฐอาหรับเอมิเรตส์	168
5	ลาฮอร์, ปากีสถาน	164
18	กรุงเทพฯ, ไทย	108

ที่มา : เว็บไซต์ Iqair.com รายงานคุณภาพอากาศจากทั่วโลก

Total deaths attributable to AAP in 2016, by region



AAP: Ambient air pollution; Afr: Africa; Amr: America; Emr: Eastern Mediterranean; Eur: Europe; Sear: South-East Asia, Wpr: Western Pacific; LMIC: Low- and middle-income; HIC: High-income.

Ambient air pollution : a global assessment of exposure and burden of diseas
WHO Geneva (2016)

Air pollutants

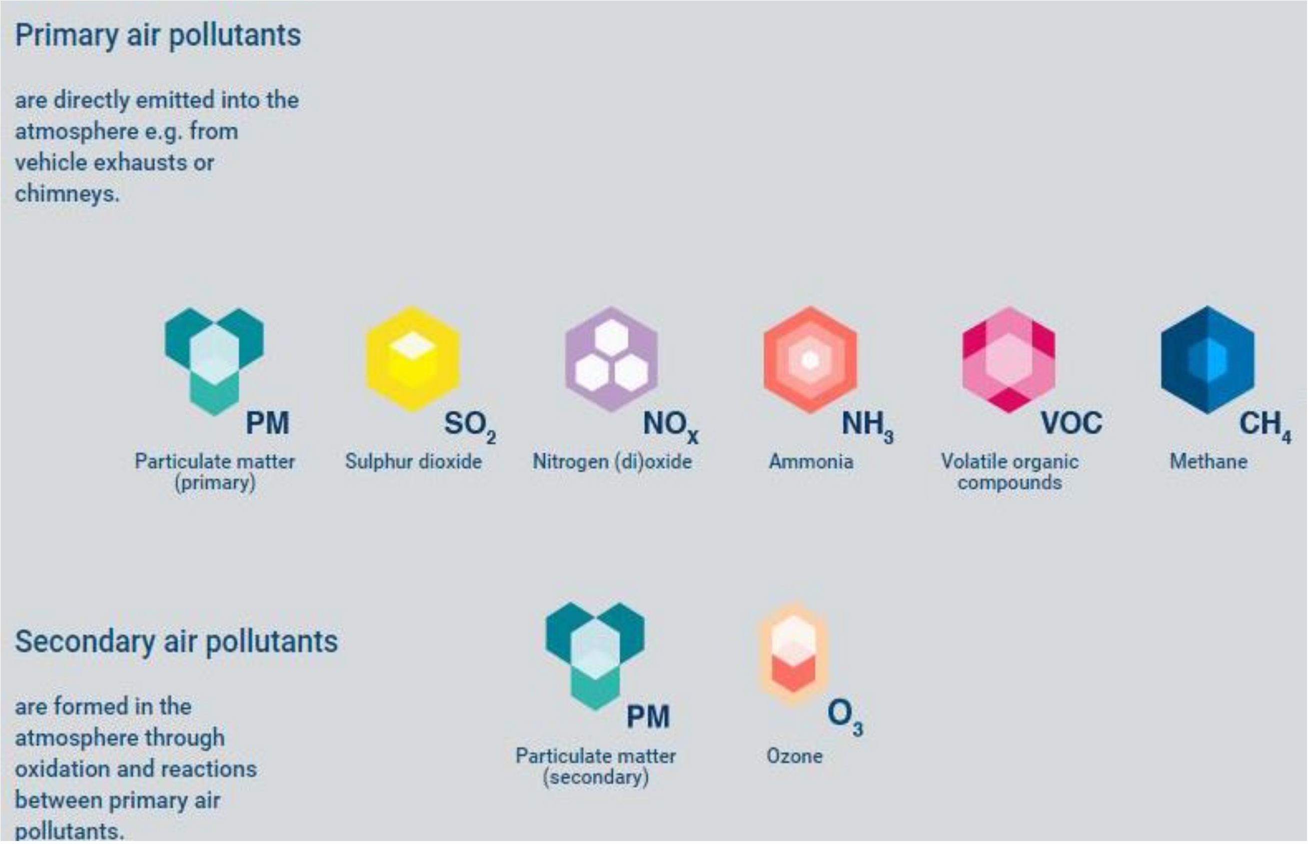
Main 6 air pollutants

Particulate matter

- Particulate matter (PM2.5)
- Particulate matter (PM10)

Gaseous pollutants

- Nitrogen oxides (NO_x)
- Sulfur dioxide (SO_2)
- Ozone (O_3)
- Carbon monoxide (CO)



Primary air pollutants

are directly emitted into the atmosphere e.g. from vehicle exhausts or chimneys.

Secondary air pollutants

are formed in the atmosphere through oxidation and reactions between primary air pollutants.

ราชวิทยาลัยอายุรแพทย์แห่งประเทศไทย
The Royal College of Physicians of Thailand (RCPT)

RCPT initiative research

Approved 2021

Project overview: Collaboration

ราชวิทยาลัยอายุรแพทย์แห่งประเทศไทย
The Royal College of Physicians of Thailand (RCPT)

กรมควบคุมมลพิษ
POLLUTION CONTROL DEPARTMENT
กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม

กรุงเทพมหานคร

สำนักสิ่งแวดล้อม

HDC

กระทรวงสาธารณสุข
MINISTRY OF PUBLIC HEALTH

GISTDA
สำนักงานพัฒนาเทคโนโลยีอวกาศและภูมิสารสนเทศ (องค์การมหาชน)

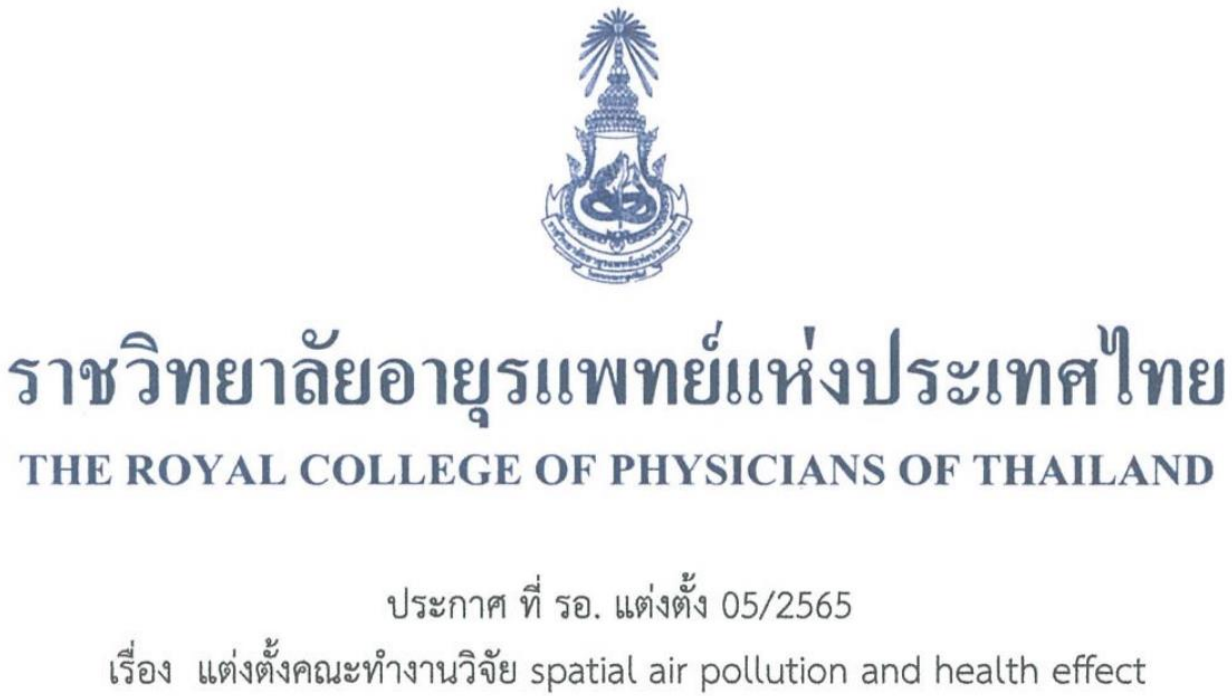
สำนักอนามัย
กรุงเทพมหานคร

สำนักการแพทย์
กรุงเทพมหานคร

สมาคม โรคสมองเสื่อมประเทศไทย
The Dementia Association of Thailand

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Project overview: collaboration



ราชวิทยาลัยอายุรแพทย์แห่งประเทศไทย

THE ROYAL COLLEGE OF PHYSICIANS OF THAILAND

ประกาศ ที่ รอ. แต่งตั้ง 05/2565

เรื่อง แต่งตั้งคณะทำงานวิจัย spatial air pollution and health effect

Project overview: Objective

To explore the association between major air pollutants and non-communicable diseases (NCDs) including death

- Longitudinal data analyses
  - Seasonal changes
  - All provinces across Thailand
- Covers various health outcomes
  - Defined by the largest government database
  - Based on both short-term and long-term exposure
- Covariates

Study theme

Study design: Retrospective population-based association study

Study setting: All 77 provinces across Thailand using national databases

Study period: 2002 to 2020

Exposure: **Air pollutants**

→ Particulate matters (PM2.5 and PM10)

→ Gaseous pollutants (NO_2, SO_2, CO, O_3)

Air quality index (AQI)

Outcome: **Non-communicable diseases (NCDs)**

→ Acute disease or Acute attack of Chronic disease

→ Chronic disease

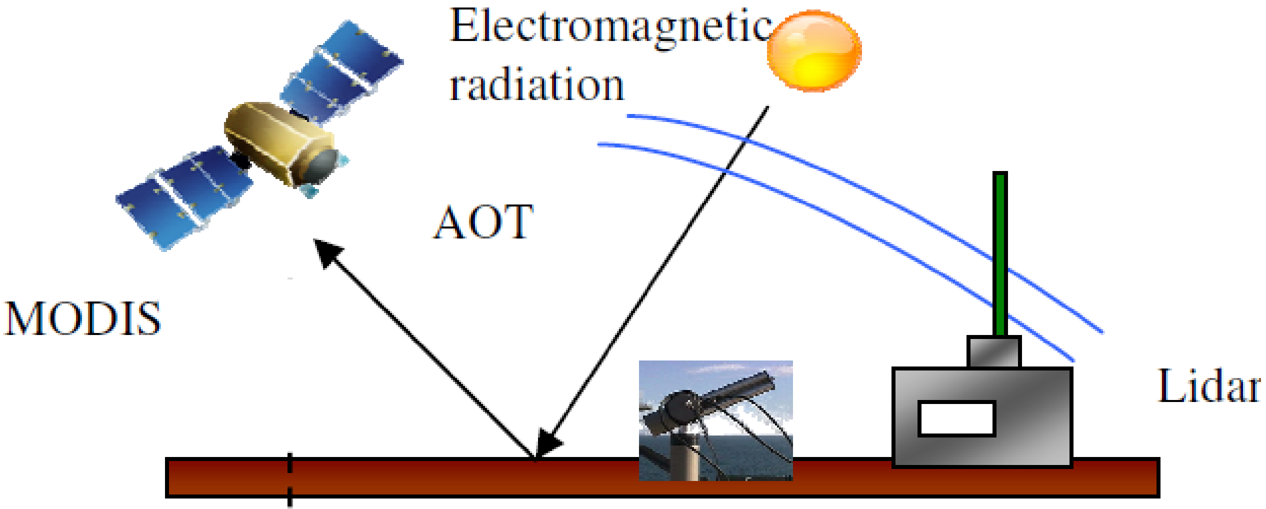
Collected data sources

Sources
Air pollutants
1. PCD (สถานีฯ กรมควบคุมมลพิษ)
2. GISTDA
3. สำนักสิ่งแวดล้อม กรุงเทพมหานคร
Outcomes
สปสช

Sources of PM2.5 data

- **Monitoring station**
  - **Pollution Control Department** of Thailand and of the Bangkok Metropolitan Administration (PCD)
  - **Department of Environment** of the Bangkok Metropolitan Administration (DOE)
- **Estimation from using satellite measurements**
  - Geo-Informatics and Space Technology Development Agency **(GISTDA)** of Thailand

กรมควบคุมมลพิษ, 2562

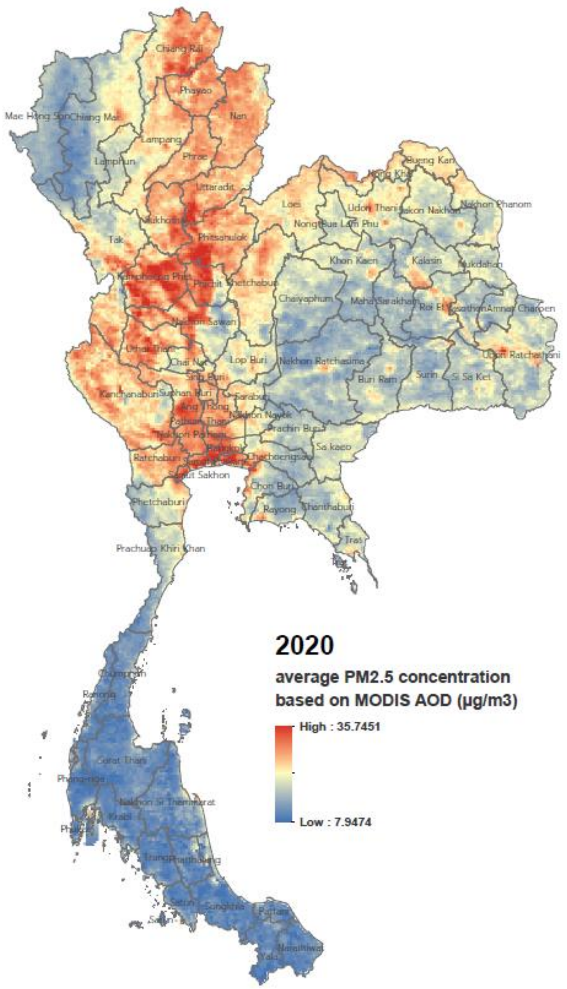


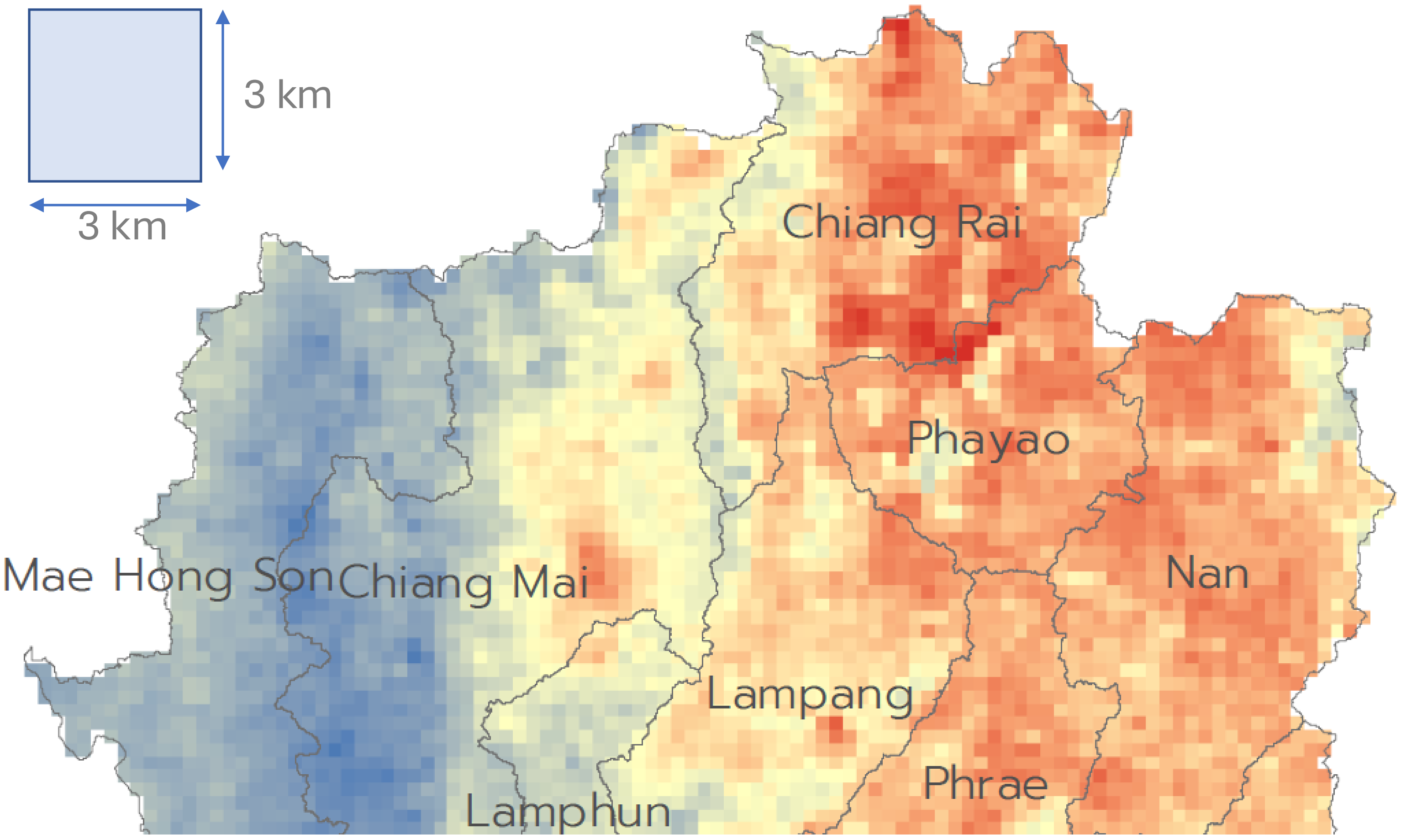
Hadjimitsis D. G. (2013)

Estimating Air pollutants

from satellite measurements

Grid size of our data: 3x3 km^2 to 10x10 km^2





Comparison

Monitoring station VS Measures from satellite

Monitoring Station

- Specific area measurement By station

  (May be considered as a point, but it is not a point measurement)
- Actual measurement From monitoring station
- Data rich in Bangkok
- 2017-2025

Measures from satellite

- Area measurement

  (by subdistrict, district, province, health region)
- Estimated data From Satellite measurements
- Cover all regions with 1x1 to 10x10 m^2 resolution
- Up to 21 years

  (Particulate matters: 2002-2025

  Gaseous pollutants: 2019-2025)

RCPT initiative research as of March ’26

Research area	Projects	Status
Mental disorder	Association between Wildfire area and PM2.5 levels on the Prevalence of Mental disorders in Thailand	Published: Environmental Challenges (Q1)
Cardiovascular	Fine Particulate Matter Exposure and Risk of Major Adverse Cardiac and Cerebrovascular Events (MACCE) in Post-Percutaneous Coronary Intervention (PCI) Patients: A Thai PCI Registry-Based Cohort Study	Published: Global Heart (Q1)
Respiratory	PM2.5 and COPD exacerbation: Aggregated data analysis	Ongoing Finalizing the analyses *Dr. Tint Lwin Win*
Respiratory	PM2.5 and COPD exacerbation: Individual data analysis with machine learning	
Measurement	Agreement among measures from monitoring station and satellite products	Finalize the analyses
Cancer	PM2.5 and head & neck and lung cancer	Ongoing *Dr. Tint Lwin Win*
Neurology	PM2.5 and Dementia	Ongoing *Dr. Htun Teza*

Association between Ambient PM 2.5 and Dementia in Thailand

Htun Teza B.D.S. M.Sc., Pawin Numthavaj M.D. Ph.D.

Department of Clinical Epidemiology and Biostatistics

Chavit Tunvirachaisakul, MD, PhD

Department of Psychiatry, Faculty of Medicine, Chulalongkorn University

Research sub-committee, The Dementia Association of Thailand

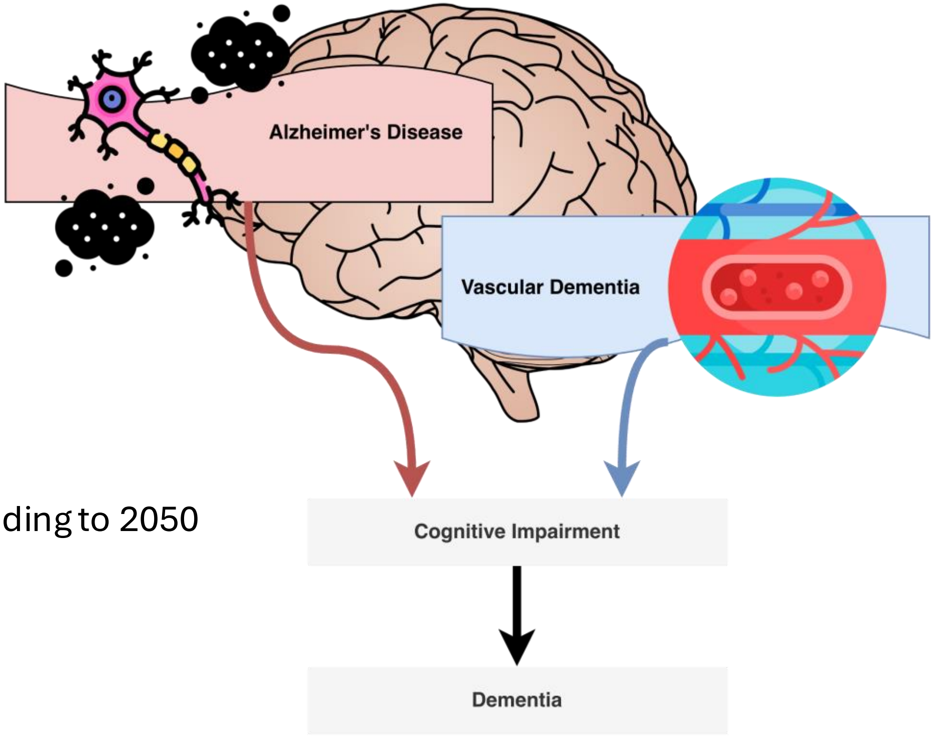
Dementia

Major Neurocognitive Disorder

- Clinical syndrome, caused by diseases that progressively destroy nerve cells
- Deterioration in cognitive functions
- Commonly, Neurodegenerative or Vascular pathology

- HICs → stabilizing or declining age-specific incidence
- LMICs → rapid increases in absolute case numbers

- Asia → worsening cardiovascular risk profiles
  → rising dementia prevalence.
- Japan → approximately 23–38% from the 1980s to the 2000s
- China → a relative annual incidence increase of around 2.9% extending to 2050 estimated



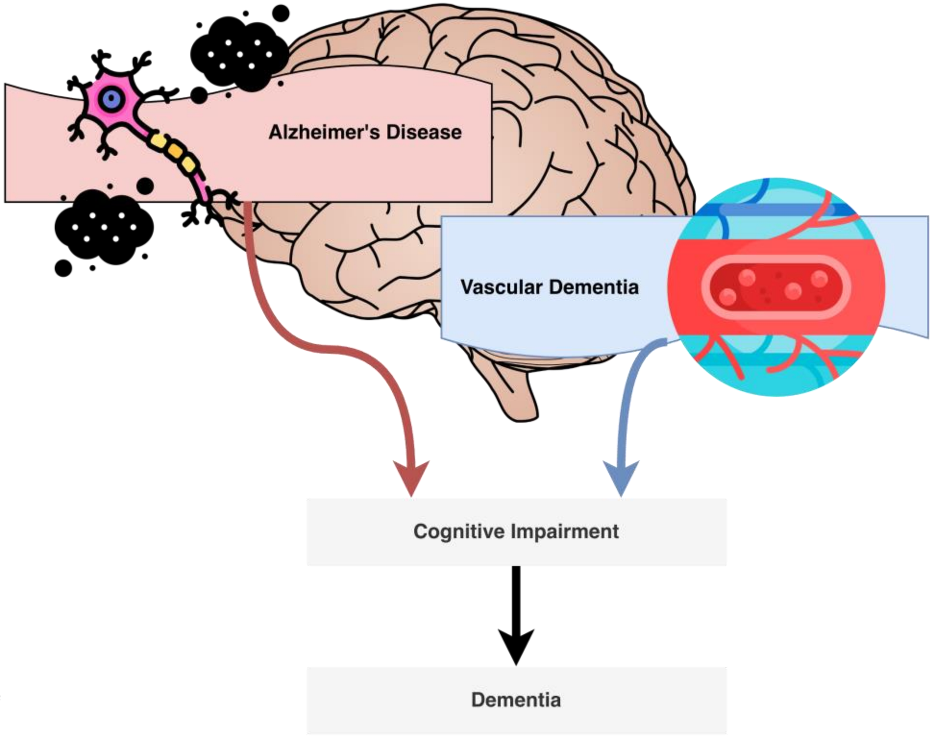
Dementia

Major Neurocognitive Disorder

- Higher long-term PM2.5 is linked to increased all-cause dementia risk,
- Pooled hazard ratios (HRs) around 1.40 (95% CI 1.23, 1.60) per 10 μg/m^3 increase reported

- PM2.5 can reach the brain via the olfactory nerve and bloodstream,
- It promotes neuroinflammation, oxidative stress, blood–brain barrier disruption, amyloid-β deposition, and tau pathology.

- PM2.5 also worsens cardiovascular disease and stroke, which themselves elevate dementia risk



Cheng, S., Jin, Y., Dou, Y., Zhao, Y., Duan, Y., Pei, H., & Lyu, P. (2022). Long-term particulate matter 2.5 exposure and dementia: a systematic review and meta-analysis. Public Health, 212, 33–41. https://doi.org/10.1016/j.puhe.2022.08.006

Real World Data

Dementia develops over many years

- Real-world evidence relies on data sources capable of long-term follow-up.

Electronic Health Records	Claims	Trials Data	Population Cohorts
Clinical Practice Research Datalink (CPRD), UK	Medicare, US	Systolic Blood Pressure Intervention Trial (SPRINT), US	Rotterdam Elderly Study, Netherlands
Secure Anonymized Information Linkage (SAIL) Databank, UK		Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), US	Leiden 85+ Study , Netherlands
Epic-based provider networks , US			Kungsholmen Project (SNAC-K), Sweden
Veterans Health Administration records , US			Newcastle 85+ Study, UK

Real World Data

Geographical Gap

- Majority of studies are done in high-income countries (HICs) in
  - North America and Europe, or
  - specific East Asian jurisdictions like Taiwan.
- These settings feature mature longitudinal infrastructures and higher hypertension control rates.

- Thailand → Middle-Income Country (MIC) landscape with worsening cardiovascular risk profiles
- A Southeast Asian or Thai dataset therefore offers critical geographic diversity

Claims Data

National Health Security Office

- Visits Data : Diagnoses and Procedures conducted

  : One record per visits

- Prescription Data : Medication prescribed for the visits

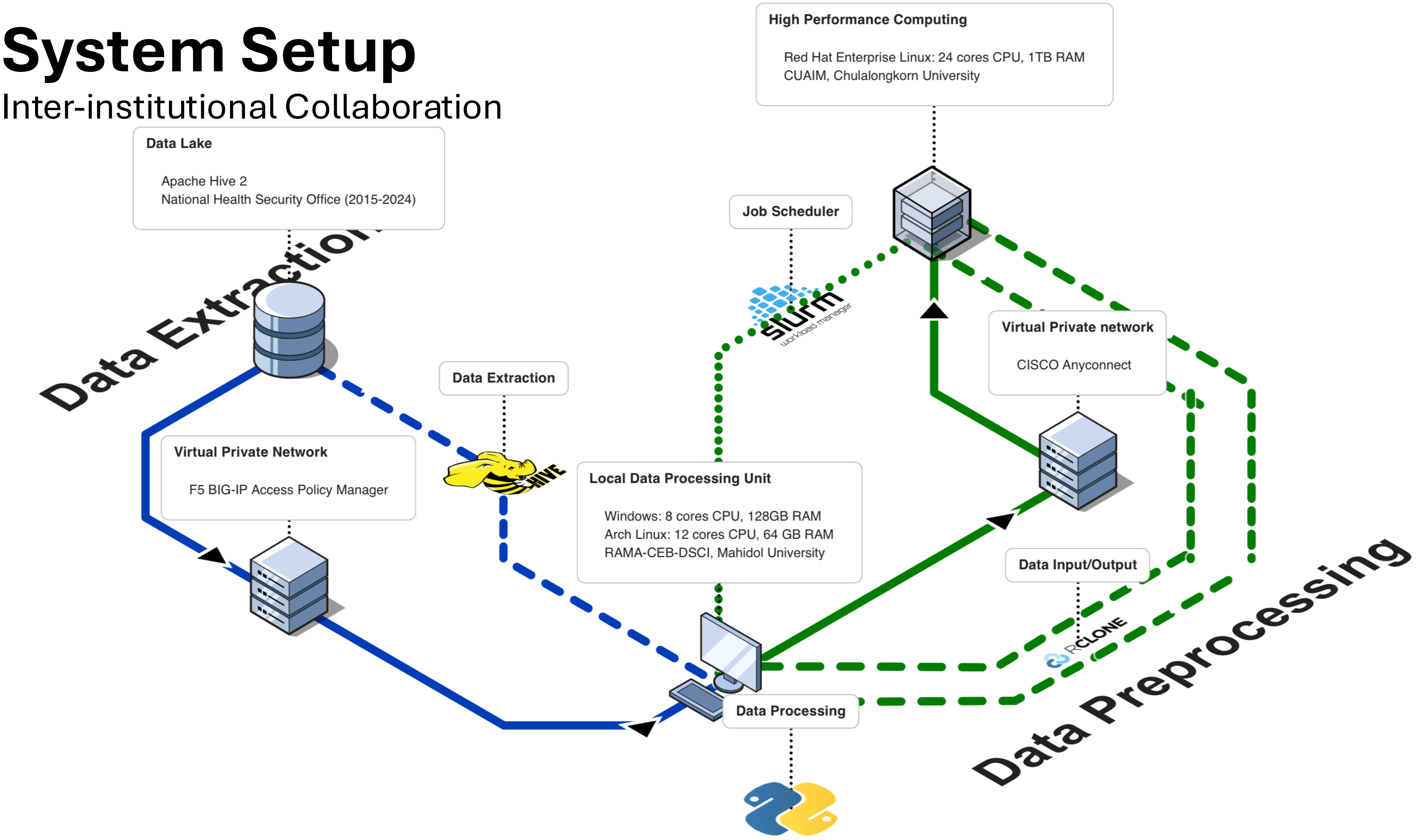
  : Multiple data entries per visit

- Claims data have unbalanced visits frequencies.
- Sick people visits more frequently while less in good health.
- The visits will be regularized into annual data.

Fiscal Year	Visits (Millions)	Prescriptions (Millions)
2559	123.5	232.0
2560	128.4	234.2
2561	122.0	230.5
2562	137.7	241.3
2563	126.7	231.9
2564	121.9	219.4
2565	125.4	227.1
2566	107.0	232.9
2567	106.2	248.1
IPD	148.3	687.0

System Setup

Inter-institutional Collaboration



High Performance Computing

Red Hat Enterprise Linux: 24 cores CPU, 1TB RAM
CUAIM, Chulalongkorn University

Data Lake

Apache Hive 2
National Health Security Office (2015-2024)

Job Scheduler

Virtual Private network

CISCO Anyconnect

Data Extraction

Virtual Private Network

F5 BIG-IP Access Policy Manager

Local Data Processing Unit

Windows: 8 cores CPU, 128GB RAM
Arch Linux: 12 cores CPU, 64 GB RAM
RAMA-CEB-DSCI, Mahidol University

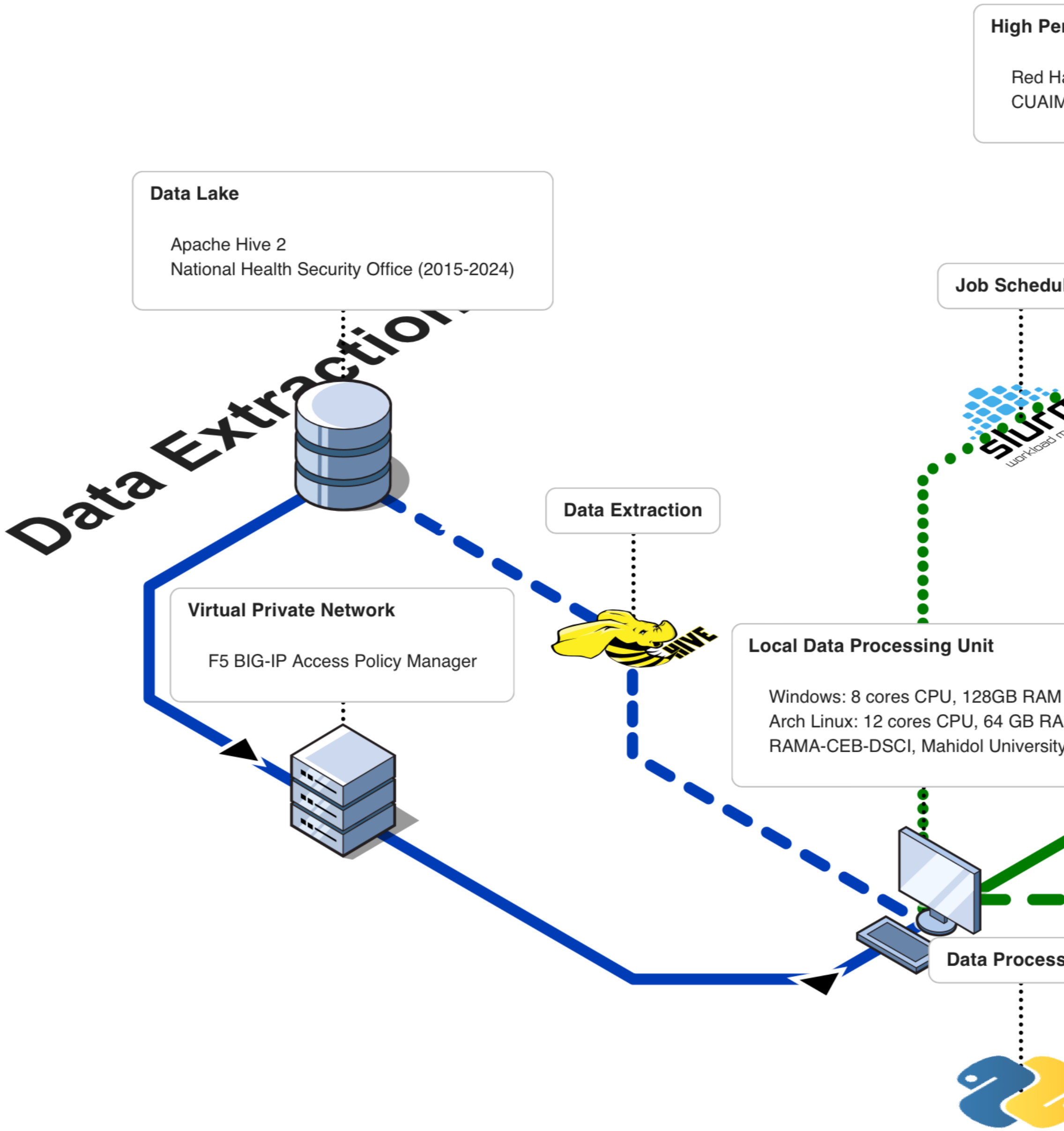
Data Input/Output

Data Processing

System Setup

Quantity (Data Scale)

- National claims data at tens to hundreds of GB per file (visit + prescription layers)
- Longitudinal coverage across multiple fiscal years → multi-terabyte total volume
- Does not fit into memory or storage of standard consumer machines
- Requires distributed storage and chunk-based processing



High Per
Red Ha
CUAIM

Data Lake
Apache Hive 2
National Health Security Office (2015-2024)

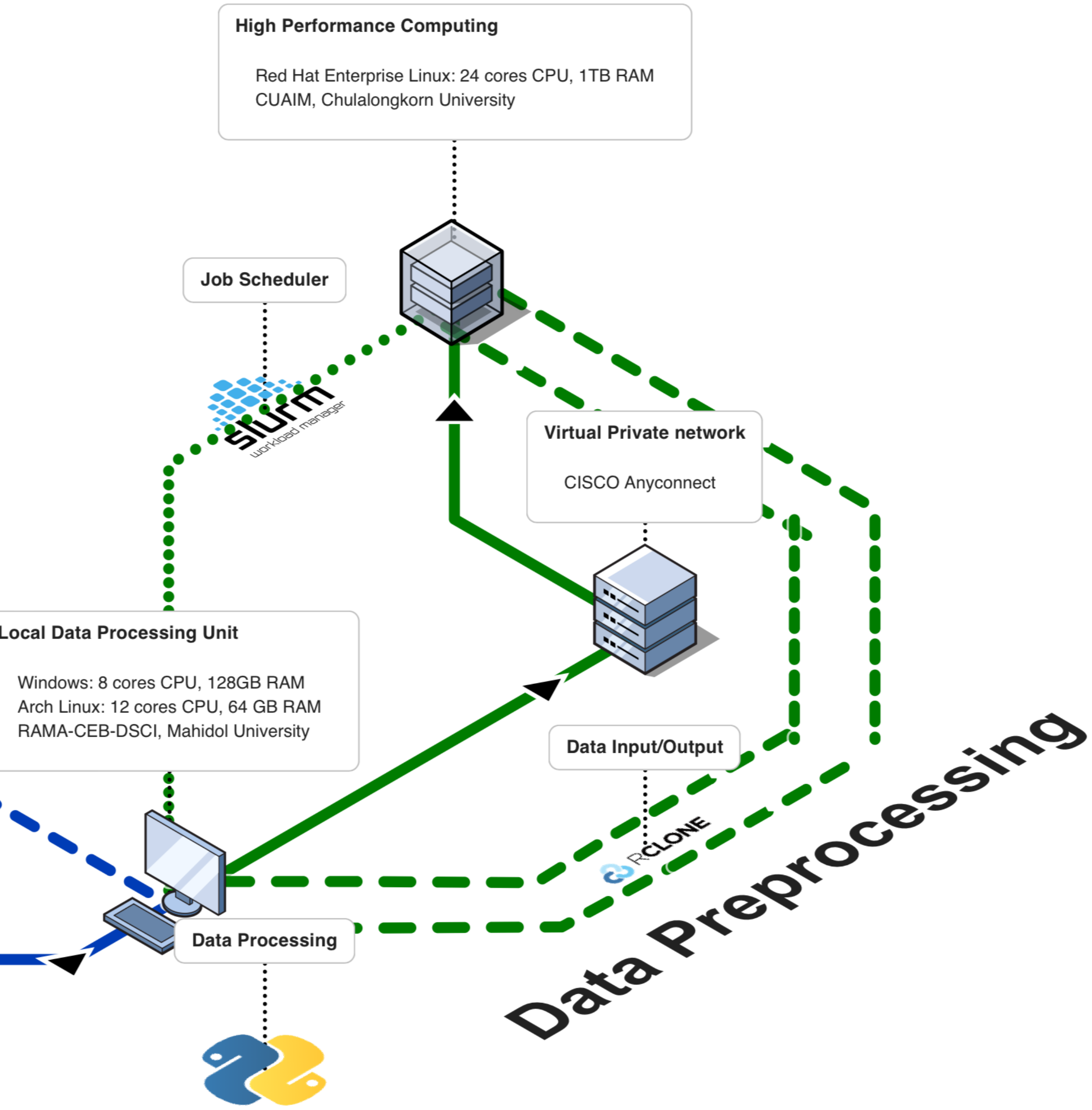
Job Schedu

Data Extraction

Virtual Private Network
F5 BIG-IP Access Policy Manager

Local Data Processing Unit
Windows: 8 cores CPU, 128GB RAM
Arch Linux: 12 cores CPU, 64 GB RA
RAMA-CEB-DSCI, Mahidol University

Data Process


High Performance Computing
Red Hat Enterprise Linux: 24 cores CPU, 1TB RAM
CUAIM, Chulalongkorn University
Job Scheduler
slurm workload manager
Virtual Private network
CISCO Anyconnect
Local Data Processing Unit
Windows: 8 cores CPU, 128GB RAM
Arch Linux: 12 cores CPU, 64 GB RAM
RAMA-CEB-DSCI, Mahidol University
Data Input/Output
RCLONE
Data Processing
Data Preprocessing

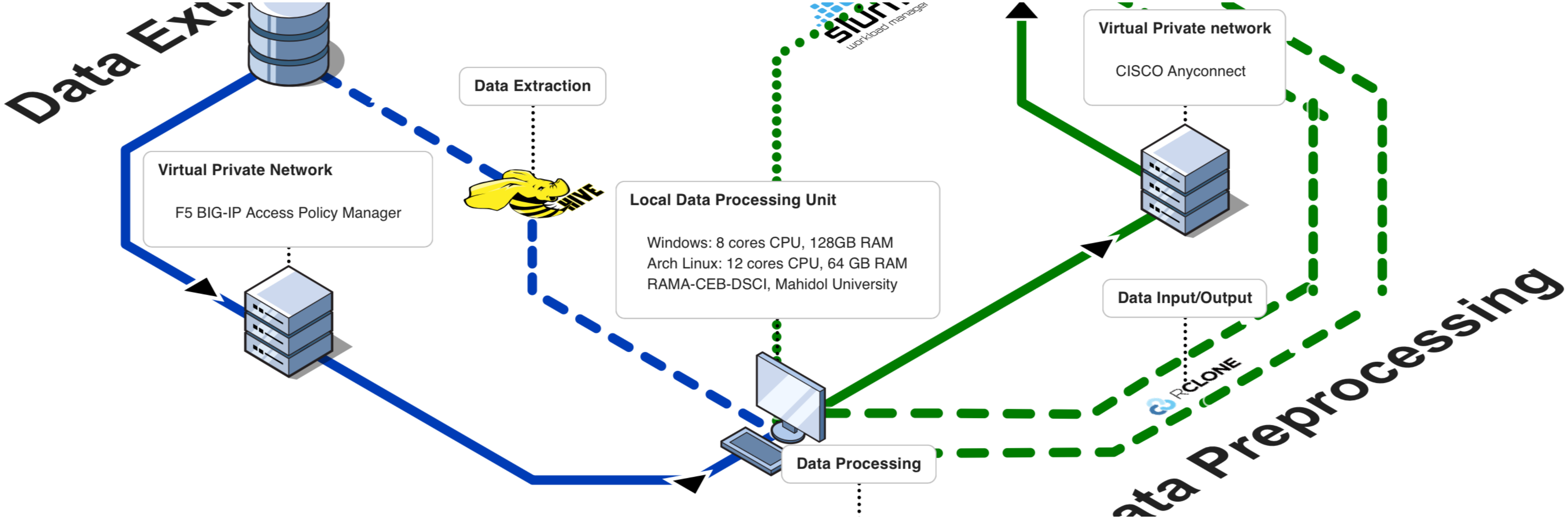
System Setup

Computation (Processing Requirements)

- Complex preprocessing:
  - Visit–prescription linkage (one-to-many structure)
  - Temporal aggregation and cohort construction
- High RAM + parallel compute required for:
  - Grouping / joins at national scale
  - Iterative cohort refinement
- Local machines (≤128 GB RAM) are insufficient for full pipeline execution
- High Performance Computing provided by Center for AI in Medicine (CU-AIM), Faculty of Medicine, Chulalongkorn University

System Setup

Security (Data Governance)



- Data contains person-level health records
- Strict access control enforced:
  - No raw data movement outside secure environment
- All transfers and computation via:
  - Private VPN tunnels
  - Controlled HPC access (CU-AIM infrastructure)
- Analysis outputs are restricted to non-identifiable aggregates

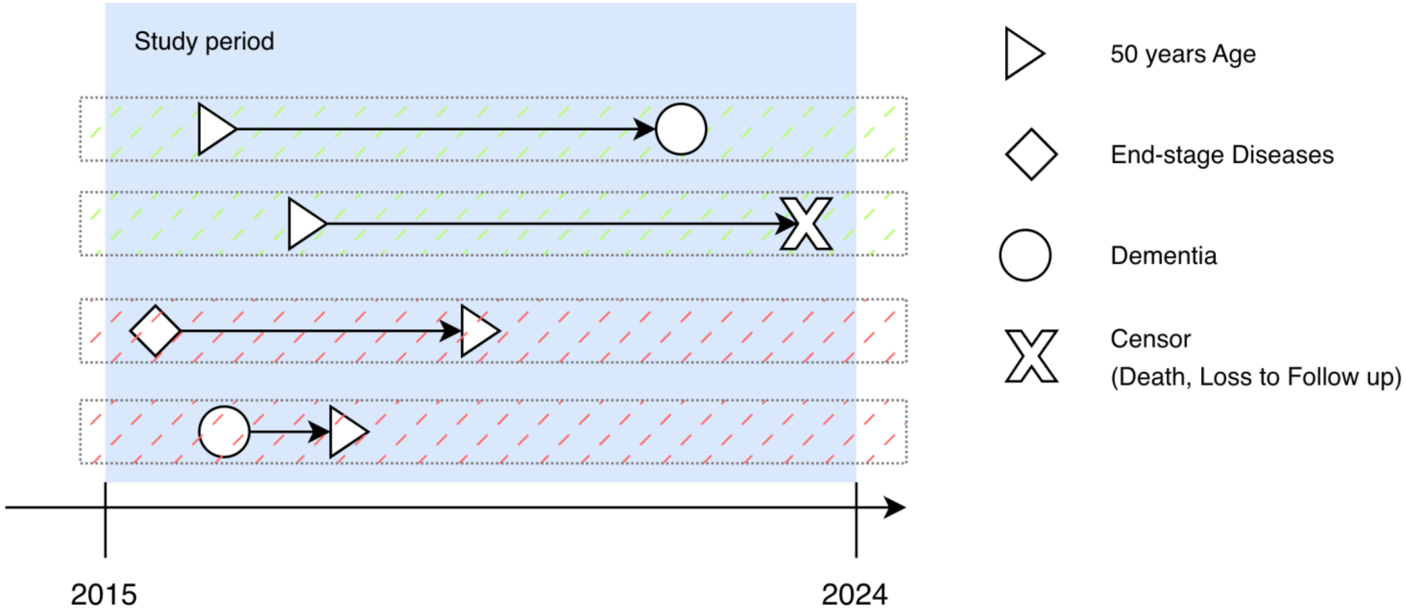
Population: Cohort

Criteria

- Adults aged 50 years or older observed in NHSO database, during the study period (2010–2024).

Exclusion Criteria

- those with end-stage chronic diseases (such as end-stage kidney failure, liver failure, or terminal cancer), and
- those diagnosed with dementia before the age of 50.



Outcome: Dementia

Criteria

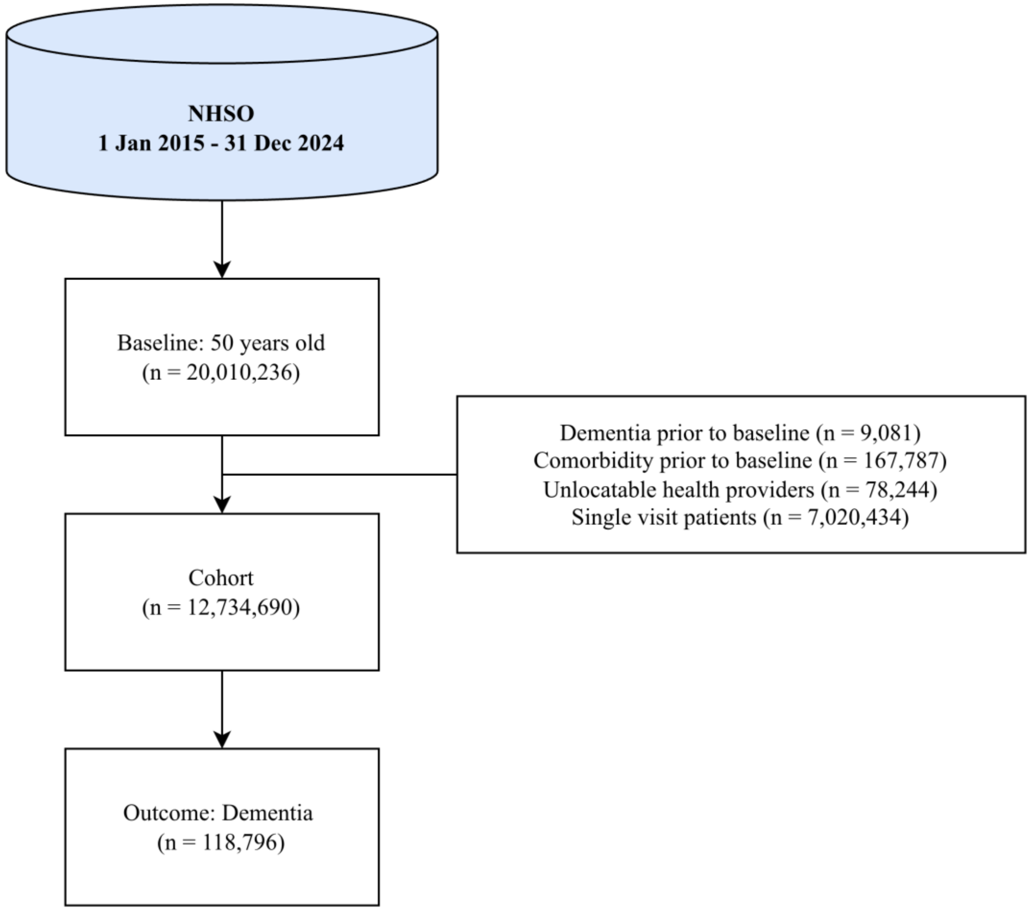
Dementia	Criteria
Alzheimer’s dementia	F000, F001, F002, F009, G300, G301, G308, G309
Vascular dementia	F010, F011, F012, F013, F018, F019
Unspecified dementia	Unspecified dementia: F03, Use of Donepezil, Rivastigmine, Galantamine, Memantine
Other dementia	Dementia with Lewy Bodies: G3183, Frontotemporal Dementia: G310, Multiple System Atrophy: G232, G233, Dementia in other diseases classified elsewhere: F02*, Progressive Supranuclear Palsy: G231, Corticobasal degeneration: G3185, HIV Dementia: B220
Mixed Dementia	Coexistence of two or more dementia pathologies recorded in the same individual

Cohort

Identification

- For 12,734,690 subjects with total 74,958,678.278 person-years
- 118,796 number of subjects develop dementia.
- Incidence rate : 1.585 (1.576, 1.594) per 1000 person-year.

	Follow-up period (years)	Observations
Mean	5.89 (2.91)	5.49 (3.04)
Median	6.70	5.00
Range	0.003 – 9.94	2.00 – 12.00
IQR	3.38-8.67	3.00 – 9.00



NHSO
1 Jan 2015 - 31 Dec 2024

Baseline: 50 years old
(n = 20,010,236)

Dementia prior to baseline (n = 9,081)
Comorbidity prior to baseline (n = 167,787)
Unlocatable health providers (n = 78,244)
Single visit patients (n = 7,020,434)

Cohort
(n = 12,734,690)

Outcome: Dementia
(n = 118,796)

Dementia

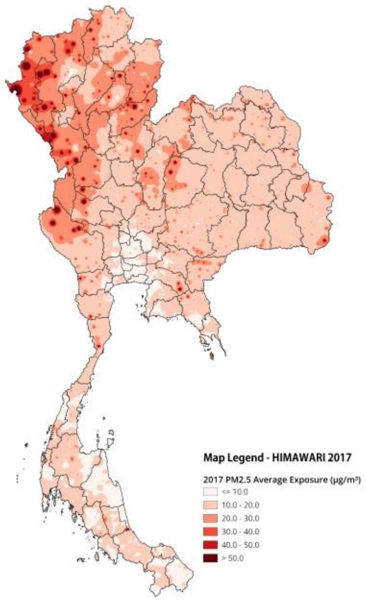
Criteria

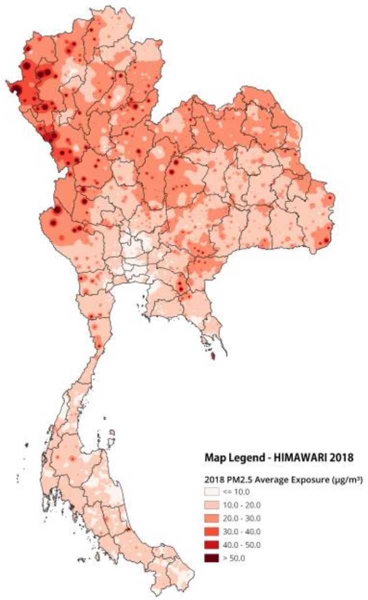
Dementia	**N**	**Cohort =** 12,734,690	**Dementia** = 118,796
Alzheimer’s disease	49,163	0.39 %	41.38 %
Vascular dementia	11,188	0.09 %	9.42 %
Unspecified dementia	28,271	0.22 %	23.8 %
Other dementia	499	0.004 %	0.43 %
Mixed Dementia	29,675	0.23 %	24.98 %

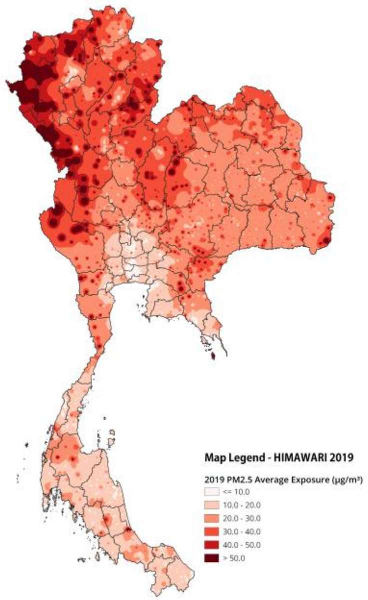
Exposure: PM2.5 Data

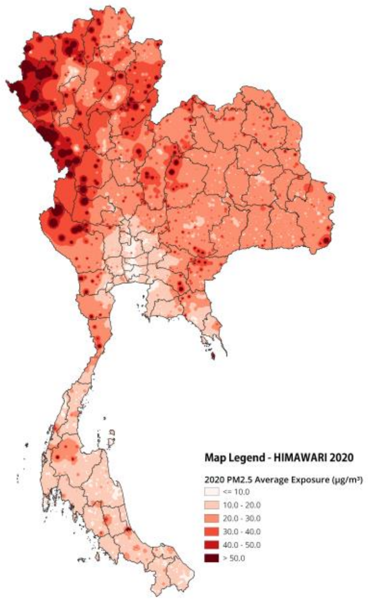
Himawari-8-AHI Aerosol Optical Depth (AOD) Level-3 data

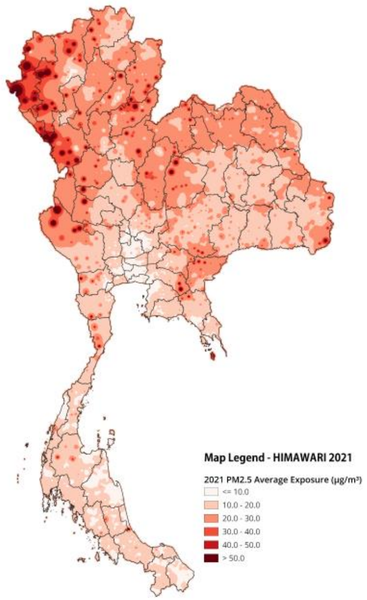
- Preprocessed for hourly data for PM 2.5 by GISTDA at Subdistrict (tambon/khwaeng) level
- (Daily) Total PM 2.5 and (Daily) Average PM 2.5 for the area (Raw and Weighted for available measurements)
- Further processed for Annual Average and Annual Maximum at district level (amphoe/khet)


Map Legend - HIMAWARI 2017
2017 PM2.5 Average Exposure (µg/m³)
<= 10.0
10.0 - 20.0
20.0 - 30.0
30.0 - 40.0
40.0 - 50.0
> 50.0


Map Legend - HIMAWARI 2018
2018 PM2.5 Average Exposure (µg/m³)
<= 10.0
10.0 - 20.0
20.0 - 30.0
30.0 - 40.0
40.0 - 50.0
> 50.0


Map Legend - HIMAWARI 2019
2019 PM2.5 Average Exposure (µg/m³)
<= 10.0
10.0 - 20.0
20.0 - 30.0
30.0 - 40.0
40.0 - 50.0
> 50.0


Map Legend - HIMAWARI 2020
2020 PM2.5 Average Exposure (µg/m³)
<= 10.0
10.0 - 20.0
20.0 - 30.0
30.0 - 40.0
40.0 - 50.0
> 50.0


Map Legend - HIMAWARI 2021
2021 PM2.5 Average Exposure (µg/m³)
<= 10.0
10.0 - 20.0
20.0 - 30.0
30.0 - 40.0
40.0 - 50.0
> 50.0

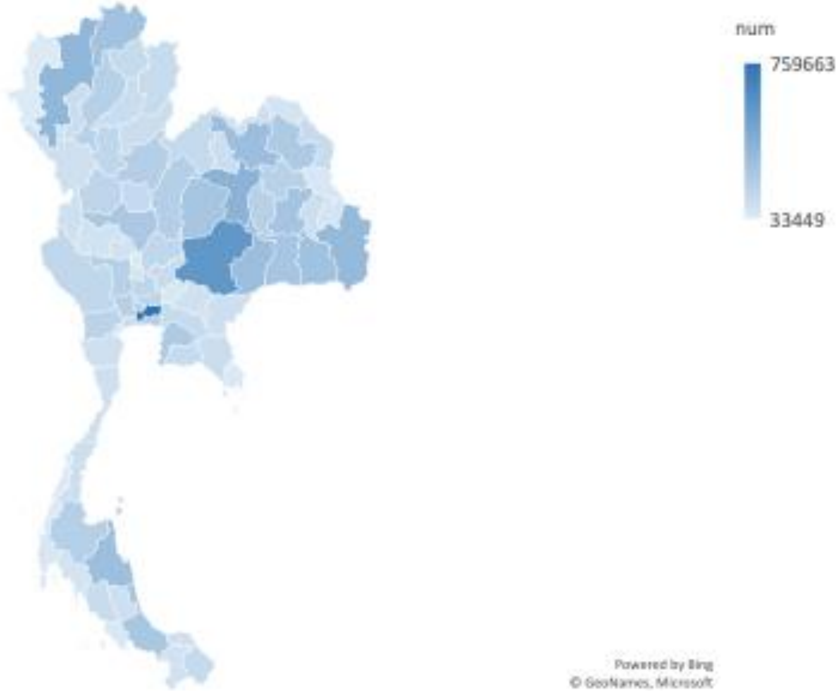
Location Data

National Health Security Office

- Patient residential information were not made available with investigators due to PDPA.

From each visit or prescription data,

- The location of hospital where the patient's health coverage is assigned to, is inferred as patient's residential location.
- It can be mapped to province (Changwat), district (amphoe/khet) and subdistrict (tambon/khwaeng) levels.



Code	No. Subjects	Name	Province	District	Subdistrict	NHSO ZONE
10685	51,784	รพ.สมุทรปราการ	สมุทรปราการ	เมืองสมุทรปราการ	ปากน้ำ	6

Analysis: Cox Proportional Hazards Model

- Individual level – Demographics and Cardiovascular profile
- District level – PM2.5 measurements
- Time level – Time varying covariates at annual level
- PM 2.5 at current time, or cumulative exposure (since 50 years of age) until current process.

Hashed ID	Visit Date	Gender	Age	Atrial Fibrillation	PM 2.5 (Annual Max)	Cumulative PM 2.5 (Annual Max)	Row Type
A	2015 – 10 – 08	F	50	0	102.08	102.08	Index
A	2016 – 09 – 27	F	51	1	234.14	366.22	Follow up
A	2017 – 09 – 27	F	52	1	182.99	519.21	Follow up
A	2018 – 09 – 25	F	53	1	188.09	707.30	Follow up
A	2019 – 08 – 13	F	54	1	275.69	982.99	Follow up
A	2020 – 09 – 17	F	55	1	277.11	12060.10	Dementia

[Preliminary] Univariate Analysis

PM 2.5 Data

- [Preliminary Analysis]
- Random effects of geographical difference is estimated by random intercept at Amphoe level.
- Random intercepts have median variance of 0.2435 (IQR 0.2418 – 0.2498; range 0.2398 – 0.2542).

Annual Aggregation		At current process			Cumulative until current process		
PM 2.5		**Univariate HR**	**P Value**	**Time (Hours)**	**Univariate HR**	**P Value**	**Time (Hours)**
Standard PM2.5 Measure	Annual Average	0.9936 (0.9929,0.9944)	< 0.0001	0.74	0.9974 (0.9973,0.9975)	< 0.0001	0.76
	Annual Maximum	0.9997 (0.9997,0.9997)	< 0.0001	0.86	0.9998 (0.9998,0.9998)	< 0.0001	0.9
Average PM2.5	Annual Average	0.9665 (0.9636,0.9694)	< 0.0001	0.75	0.988 (0.9875,0.9886)	< 0.0001	0.82
	Annual Maximum	0.9992 (0.9991,0.9992)	< 0.0001	0.9	0.9995 (0.9995,0.9995)	< 0.0001	1.02
Weighted Standard PM2.5 per hour	Annual Average	0.9971 (0.9956,0.9985)	< 0.0001	0.91	0.9975 (0.997,0.998)	< 0.0001	0.85
	Annual Maximum	0.9999 (0.9999,0.9999)	< 0.0001	0.88	0.9999 (0.9999,0.9999)	< 0.0001	0.8
Weighted Average PM2.5	Annual Average	0.9996 (0.9993,0.9999)	0.01446	0.8	0.9995 (0.9994,0.9996)	< 0.0001	0.9
	Annual Maximum	0.9999 (0.9999,0.9999)	0.00050	0.87	0.9999 (0.9999,0.9999)	< 0.0001	0.88

[Preliminary] Univariate Analysis

Demographics and Comorbidity Data

Variable		Univariate HR	P Value	Time (Hours)
Gender	Male	1		
	Female	1.0678 (1.0554,1.0804)	< 0.0001	0.65

Variable		Univariate HR	P Value	Time (Hours)
Stroke	None	1		
	Other	3.0222 (2.9634,3.0822)	< 0.0001	0.96
	Ischemic	3.0018 (2.9396,3.0653)	< 0.0001	0.96
Traumatic Brain Injury		2.4595 (2.3786,2.5431)	< 0.0001	1.11

- Dyslipidemia – Pure hypercholesterolemia, Pure hyperglyceridemia, Mixed
- Other Stroke – Hemorrhagic Stroke, Stroke sequelae of cerebrovascular disease, Other cerebrovascular conditions, Stroke non-specified

Variable	Univariate HR	P Value	Time (Hours)
Age, years	1.0787 (1.0782,1.0793)	< 0.0001	0.9
Hypertension	0.9784 (0.9664,0.9905)	0.00050	0.66
Type 2 Diabetes	0.9227 (0.9104,0.9351)	< 0.0001	0.77
Chronic Kidney Disease	1.2444 (1.2247,1.2645)	< 0.0001	0.75
Coronary Artery Disease	1.424 (1.3921,1.4566)	< 0.0001	1.15
Peripheral Vascular Disease	1.0304 (1.0072,1.0541)	0.00975	0.65
Heart Failure	1.4854 (1.444,1.5281)	< 0.0001	1.33
Atrial Fibrillation	1.6373 (1.5899,1.686)	< 0.0001	0.84
Dyslipidemia	0.9619 (0.9496,0.9744)	< 0.0001	0.65
Obesity	0.5294 (0.4919,0.5697)	< 0.0001	0.93

[Preliminary] Univariate Analysis

Demographics and Comorbidity Data

Variable	Univariate HR	P Value	Time (Hours)
COPD	1.2501 (1.2134,1.2879)	< 0.0001	0.64
Rhinitis	0.9771 (0.9485,1.0067)	0.12890	0.79
Psychiatric Disorders			
Anxiety	2.1263 (2.0745,2.1793)	< 0.0001	0.9
Bipolar affective disorder	4.7018 (4.3799,5.0474)	< 0.0001	0.8
Depression	3.4452 (3.3699,3.5221)	< 0.0001	0.85
Schizophrenia	3.5608 (3.4537,3.6713)	< 0.0001	0.7

- Rhinitis – Vasomotor and allergic rhinitis, Chronic rhinitis, nasopharyngitis and pharyngitis
- Audio-Visual Impairment - Hearing impairment, Visual impairment (Blindness, binocular; Severe, binocular; Moderate, binocular; Mild, binocular)

Variable	Univariate HR	P Value	Time (Hours)
Audio-Visual Impairment	1.3955 (1.3687,1.4228)	< 0.0001	0.71
Substance Abuse (Overall)	0.8234 (0.8035,0.8438)	< 0.0001	0.97
Sleep Disorders (overall)	1.7501 (1.7006,1.8009)	< 0.0001	0.73
STD (HIV + Syphilis)	0.4258 (0.3895,0.4656)	< 0.0001	0.82

- Substance Abuse – Alcohol, Opioid, Cannabis, Sedatives or hypnotics, Cocaine + Other stimulants, Hallucinogen, Tobacco, Volatile solvent, Other substances
- Sleep Disorders – Insomnias, Sleep-wake cycles, Sleep apnea

Continuing Works

- Paradoxical observations are being addressed
- Careful modelling is required for PM 2.5 index.
- Additional feature specifications
    - Hypertension, Diabetes : (yes/ no) → (yes with medication/ yes without medication/ no)
- Confounding features should be taken into account, such as cardiovascular conditions, dyslipidemia and obesity.
- Several features might be under-coded, such as substance abuse and obesity.

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

CEB

Association between ambient $PM_{2.5}$ exposure and the occurrence of acute exacerbations of chronic obstructive pulmonary disease (COPD)

Tint Lwin Win M.B.B.S., Pawin Numthavaj M.D. Ph.D.

Department of Clinical Epidemiology and Biostatistics

Faculty of Medicine Ramathibodi Hospital, Mahidol University

Outline

- COPD cohort identification
- Outcome identification
- Data preparation
- Data analysis models

COPD cohort identification criteria

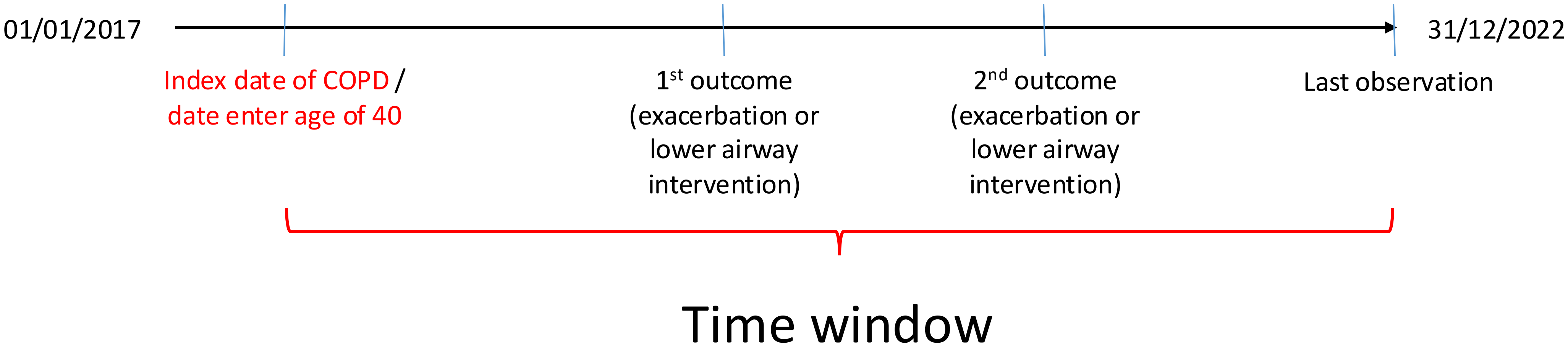
1. Diagnosed as **COPD**:
    1. **ICD10 diagnosis codes:** J44 (J440, J441, J442, J448, J449)
    2. **Condition:** Diagnosis at any point in time.
2. Confirmed by **lower airway medication**:
    1. **Condition:** Prescribed at any point in time.
    2. **Clarification:** We use the list of medications that qualify as "lower airway medication" (mucolytics, expectorants, bronchodilators, corticosteroids, etc.).

Outcome identification

1. **COPD exacerbation**
   - Identified by ICD10: J441
2. **Lower airway intervention (with lower airway medication):** Identified by ICD9
   - **Non-invasive mechanical ventilation**: 93.90
   - **Nebulizer**: 93.94
   - **Endotracheal tube**: 96.04
   - **Temporary tracheostomy**: 31.1
   - **Other invasive mechanical ventilation**: 96.7

Study timeline for COPD cohort

Cohort of COPD patients with repeated outcomes of exacerbation and lower airway intervention over time



01/01/2017

31/12/2022

Index date of COPD / date enter age of 40

1st outcome (exacerbation or lower airway intervention)

2nd outcome (exacerbation or lower airway intervention)

Last observation

Time window

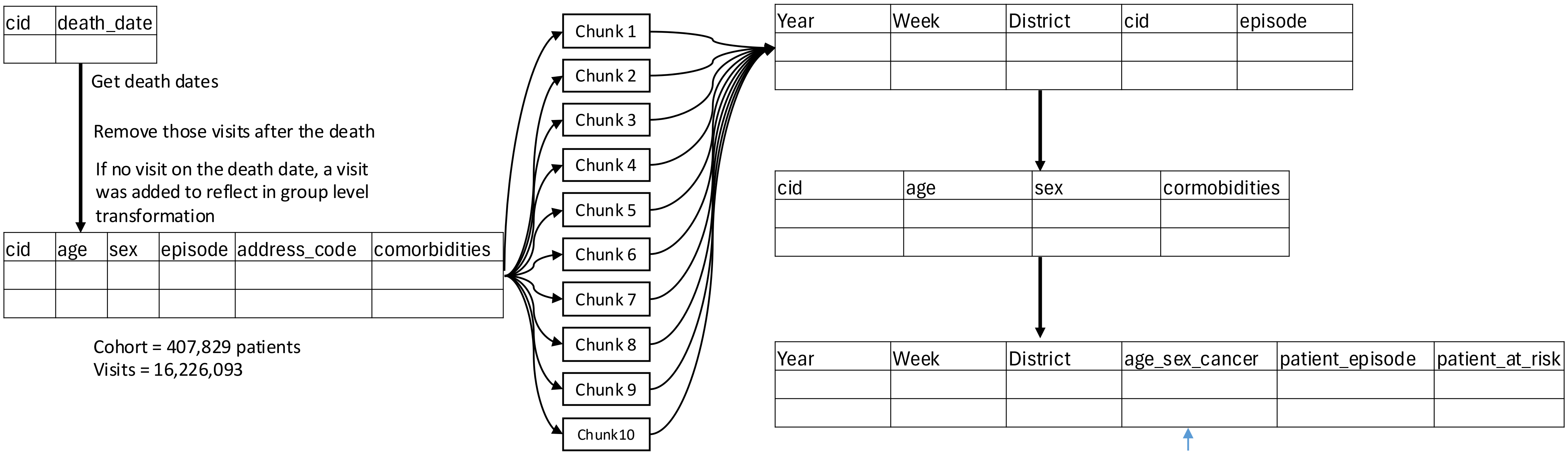
Summary statistics

Characteristics of COPD cohort		n	%
Total patient in 6 years (2017-2022)		407866	100.0
Age	Median (IQR)	68 (60 - 77)	
Sex	Male	292,870	71.8%
	Female	114,996	28.2%

Data preparation - group level

Individual level data – COPD cohort with episodes

Weekly district level data



cid	death_date

Get death dates

Remove those visits after the death

If no visit on the death date, a visit was added to reflect in group level transformation

cid	age	sex	episode	address_code	comorbidities

Cohort = 407,829 patients
Visits = 16,226,093

Each chunk takes around 8 hours for group level data, 1-2 hours for crosstabulations (age, sex, cancer)

Year	Week	District	cid	episode

cid	age	sex	cormobidities

Year	Week	District	age_sex_cancer	patient_episode	patient_at_risk

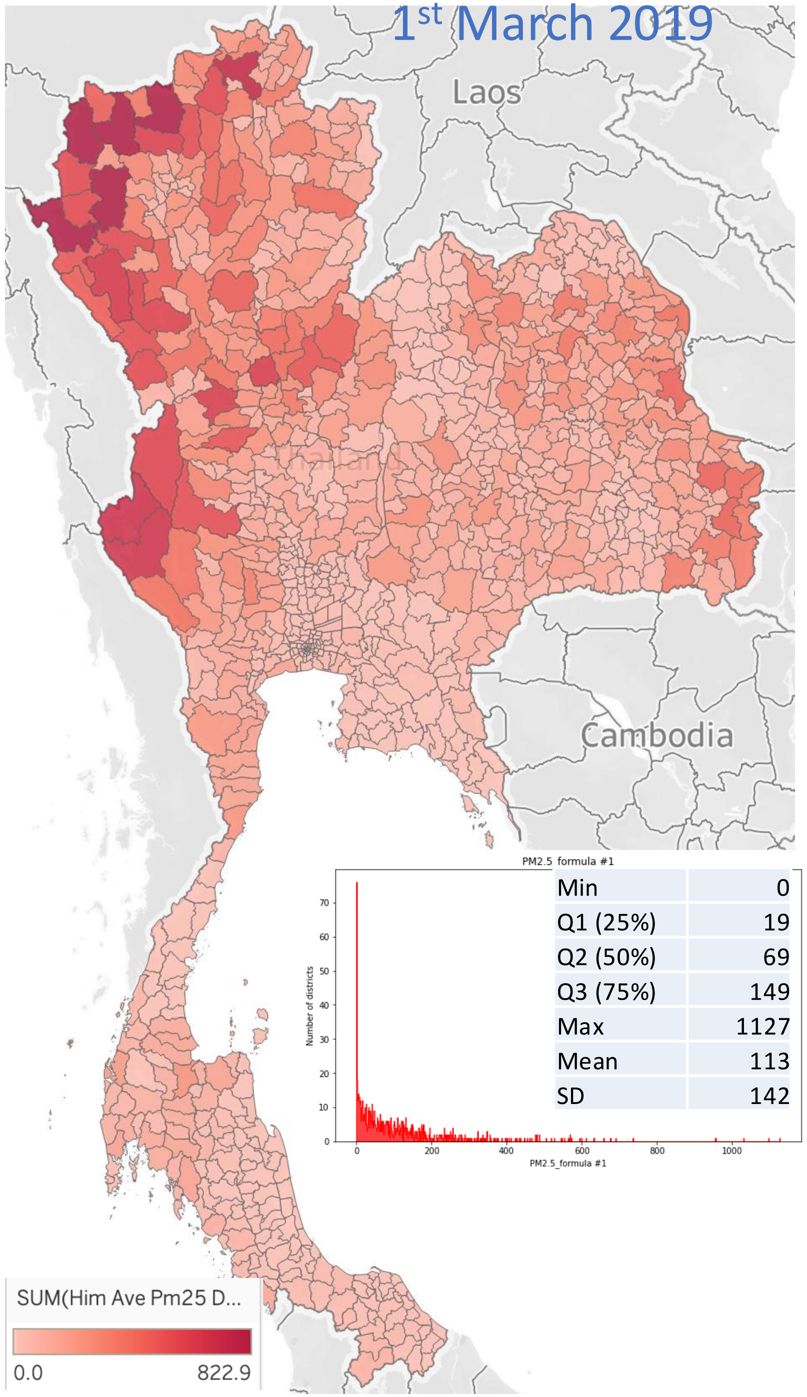
Age group = 40+, 50+, 60+ and 70+
Sex = Male / Female
Cancer-Yes / No

Week = 313
District = 927
Age_sex_cancer = 16 groups

Data preparation – individual level data

- **Rationale for Lumping:**
  - Some patients experience multiple COPD outcomes in a short period.
  - To streamline data analysis and reduce redundancy, outcome visits are grouped.
- **COPD Outcomes Lumping:**
  - **Criteria:** COPD outcomes are grouped into a single episode if they occur within 28 days of each other.

1st March 2019

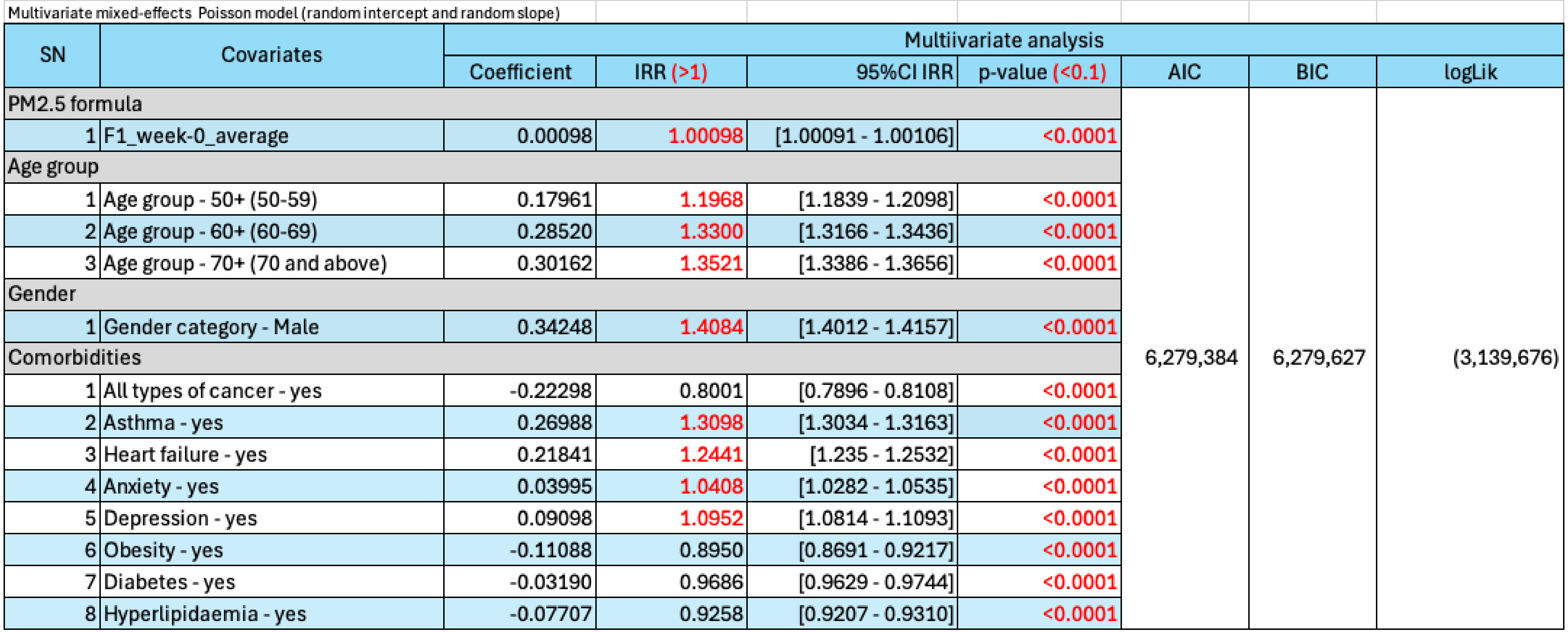


Min	0
Q1 (25%)	19
Q2 (50%)	69
Q3 (75%)	149
Max	1127
Mean	113
SD	142

Poisson analysis

- **Dependent variable** = number of patients with episode
- **Offset** = number of patients at risk
- **link** = "log”
- **Cluster** = district
- **Predictors** =PM2.5 , age group, gender, comorbidities

Multivariate Mixed-effects Poisson regression



Multivariate mixed-effects Poisson model (random intercept and random slope)

SN	Covariates	Multiivariate analysis						
		Coefficient	IRR (>1)	95%CI IRR	p-value (<0.1)	AIC	BIC	logLik
PM2.5 formula						6,279,384	6,279,627	(3,139,676)
1	F1_week-0_average	0.00098	1.00098	[1.00091 - 1.00106]	<0.0001			
Age group								
1	Age group - 50+ (50-59)	0.17961	1.1968	[1.1839 - 1.2098]	<0.0001			
2	Age group - 60+ (60-69)	0.28520	1.3300	[1.3166 - 1.3436]	<0.0001			
3	Age group - 70+ (70 and above)	0.30162	1.3521	[1.3386 - 1.3656]	<0.0001			
Gender								
1	Gender category - Male	0.34248	1.4084	[1.4012 - 1.4157]	<0.0001			
Comorbidities								
1	All types of cancer - yes	-0.22298	0.8001	[0.7896 - 0.8108]	<0.0001			
2	Asthma - yes	0.26988	1.3098	[1.3034 - 1.3163]	<0.0001			
3	Heart failure - yes	0.21841	1.2441	[1.235 - 1.2532]	<0.0001			
4	Anxiety - yes	0.03995	1.0408	[1.0282 - 1.0535]	<0.0001			
5	Depression - yes	0.09098	1.0952	[1.0814 - 1.1093]	<0.0001			
6	Obesity - yes	-0.11088	0.8950	[0.8691 - 0.9217]	<0.0001			
7	Diabetes - yes	-0.03190	0.9686	[0.9629 - 0.9744]	<0.0001			
8	Hyperlipidaemia - yes	-0.07707	0.9258	[0.9207 - 0.9310]	<0.0001			

Assessing the association between lung, head, and neck cancer incidence and exposure to PM2.5 in Thailand using traditional statistical methods and machine learning methods

Outline

- Study design and study population
- Outcome of interest identification
- Outcome of interest ascertainment
- Data preprocessing
- Data analysis plan

Study design

- This study will enroll patients who received non-communicable disease (NCD) screening in National Health Security Office (NHSO) database consisting of more than 54 million patients.
- Approximately 33 million patients has received NCD screening in NHSO database.
  - For the individual-level analysis, due to computational constraints, not all 33 million patients will be included in the analysis, and a nested case-control approach will be employed.

- Why we choose NCD screening dataset?
  - Smoking information (major confounding factor for cancers)
- What is NCD screening dataset?
  - NCD screening service provided by the Ministry of Public Health (MOPH)
  - Under the Universal Coverage Scheme (UCS)
  - For adults aged 35 years and above who have not been diagnosed with diabetes or hypertension.

Study population

- Inclusion criteria
  - Patients who have NCD screening in NHSO database.
  - Age more than or equal 35 years.
  - Patients who have more than a single observation.
- Exclusion criteria
  - Patient with cancer before the NCD screening date (index date)
  - Patient with missing smoking status in NCD screening data.

Features in NHSO data

- ICD-10 diagnosis codes.
- Demographic factors (age, sex, district of residence using hospital code)
- Behavioral risk factors (smoking status from NCD screening records)
- Comorbidities such as COPD, tuberculosis, and pulmonary fibrosis, identified through ICD-10 diagnosis code and prescription records

Outcome of interest identification

- The incidence of lung cancer, HNC retrieved using ICD-10 codes.

HNC

  - Intra-oral cancers (C00-C08)
  - Oropharyngeal cancers (C09-C11)
  - Other ill-defined sites - lip, oral cavity and pharynx (C12-C14)
  - Laryngeal cancer (C32)
  - Trachea, bronchus and lung cancer (C33 and C34)

LC

Outcome of interest ascertainment

Radiation

- Contact radiation (92.21)
- Orthovoltage radiation (92.22)
- Radioisotopic teleradiotherapy (92.23)
- Teleradiotherapy using photons (92.24)
- Teleradiotherapy using electrons (92.25)
- Teleradiotherapy of other particulate radiation (92.26)
- Implantation or insertion of radioactive elements (92.27)
- Other radiotherapeutic procedure (92.29)

Surgical procedures for lung cancer treatment

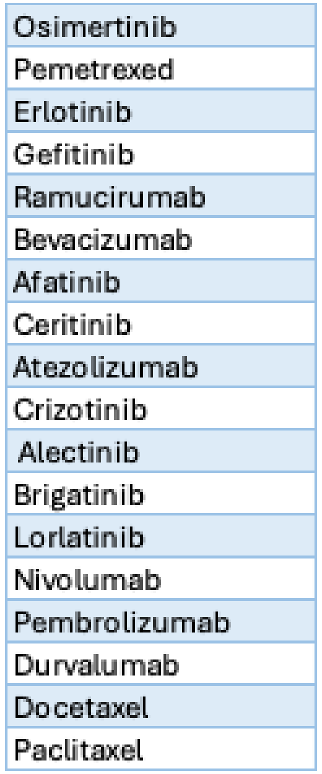
- Segmental resection of lung (32.3)
- Lobectomy of lung (32.4)
- Pneumonectomy (32.5)
- Radical dissection of thoracic structures (32.6)
- Other excision of lung (32.9)
- Incision of chest wall and pleura (34)

Surgical procedures for head and neck cancers

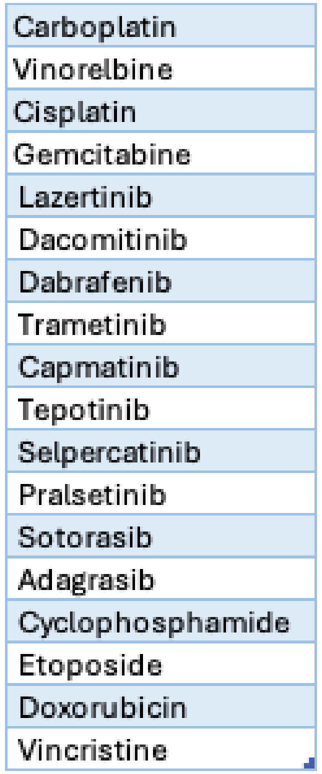
- Radical orbitomaxillectomy(16.51)
- Glossectomy (25.2-25.3)
- Wide excision or destruction of lesion or tissue of bony palate (27.32)
- Pharyngectomy (partial) (29.33)
- Laryngectomy (30.29 – 30.4)
- Radical neck dissection (40.4)
- Mandibulectomy (76.31, 76.41)
- Partial ostectomy of other facial bone (Hemimaxillectomy) (76.39)

Outcome of interest ascertainment ascertainment (cont.)

Medication for lung cancer treatment

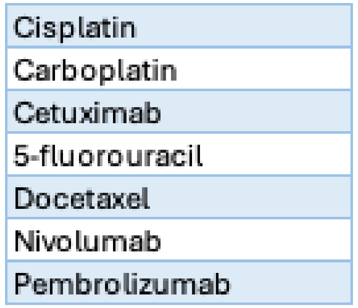


Osimertinib
Pemetrexed
Erlotinib
Gefitinib
Ramucirumab
Bevacizumab
Afatinib
Ceritinib
Atezolizumab
Crizotinib
Alectinib
Brigatinib
Lorlatinib
Nivolumab
Pembrolizumab
Durvalumab
Docetaxel
Paclitaxel



Carboplatin
Vinorelbine
Cisplatin
Gemcitabine
Lazertinib
Dacomitinib
Dabrafenib
Trametinib
Capmatinib
Tepotinib
Selpercatinib
Pralsetinib
Sotorasib
Adagrasib
Cyclophosphamide
Etoposide
Doxorubicin
Vincristine

Medication for head and neck cancers treatment



Cisplatin
Carboplatin
Cetuximab
5-fluorouracil
Docetaxel
Nivolumab
Pembrolizumab

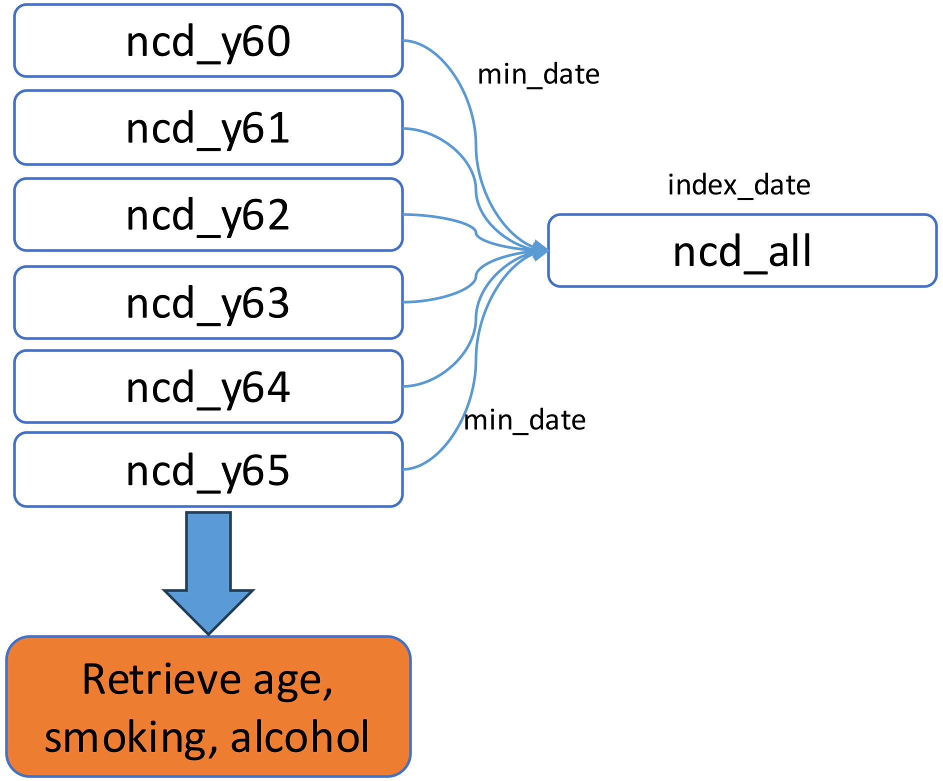
[1] Riely, G. J., Wood, D. E., Ettinger, D. S., Aisner, D. L., Akerley, W., Bauman, J. R., Bharat, A., Bruno, D. S., Chang, J. Y., Chirieac, L. R., DeCamp, M., Desai, A. P., Dilling, T. J., Dowell, J., Durm, G. A., Gettinger, S., Grotz, T. E.,Gubens, M. A., Juloori, A., … Hang, L. (2024). Non–Small Cell Lung Cancer, Version 4.2024. *JNCCN Journal of the National Comprehensive Cancer Network*, *22*(4), 249–274. https://doi.org/10.6004/jnccn.2204.0023

[2] Reungwetwattana, T., Oranratnachai, S., Puataweepong, P., Tangsujaritvijit, V., & Cherntanomwong, P. (2020). Lung Cancer in Thailand. In *Journal of Thoracic Oncology* (Vol. 15, Issue 11, pp. 1714–1721). Elsevier Inc. https://doi.org/10.1016/j.jtho.2020.04.024

[3] Guidelines for the treatment of lung cancer For use in claiming reimbursement for medical services under the National Health Security System.

NHSO data retreival

NCD screening



ncd_y60

ncd_y61

ncd_y62

ncd_y63

ncd_y64

ncd_y65

min_date

min_date

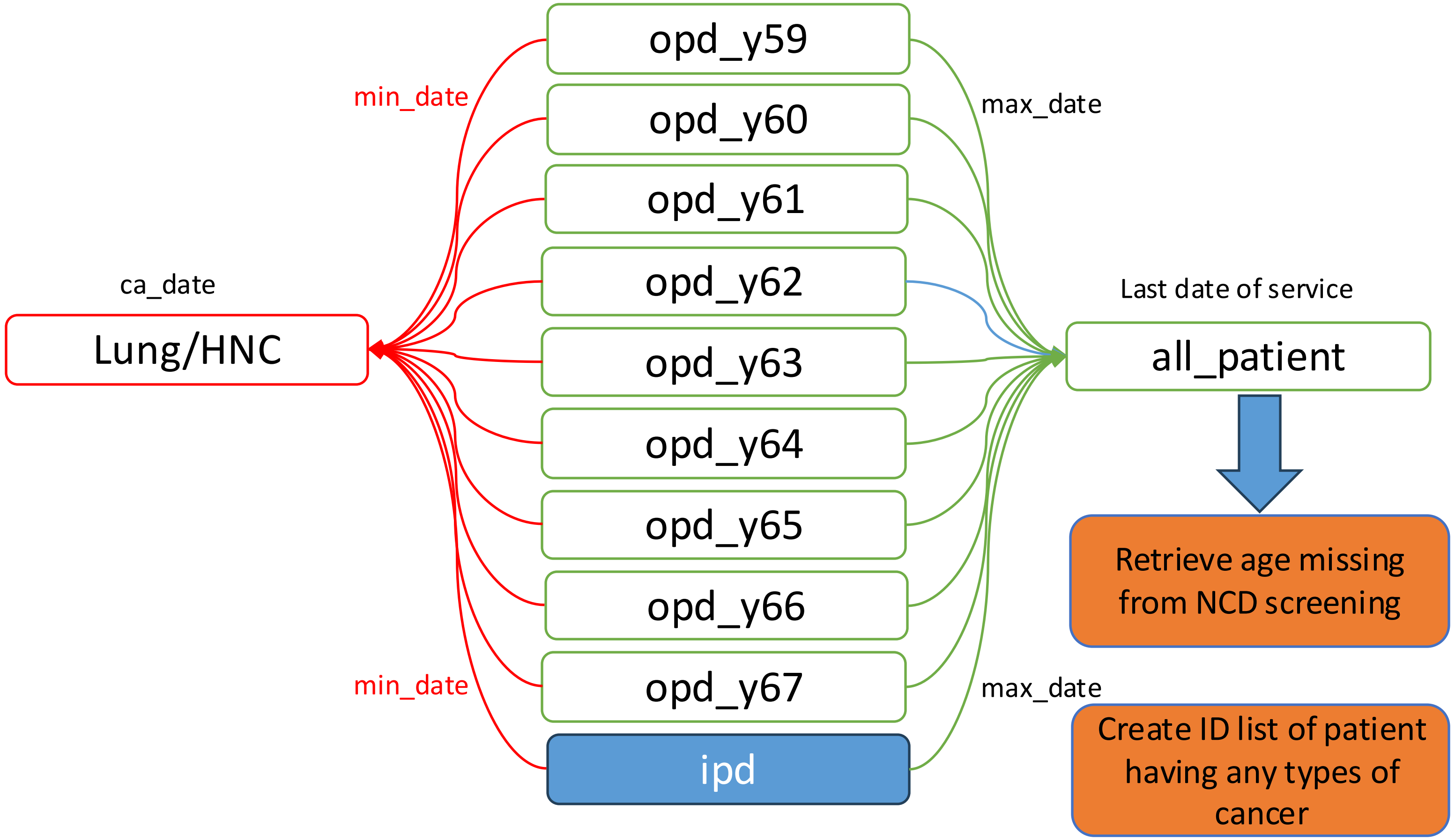
index_date

ncd_all

Retrieve age, smoking, alcohol

Inclusion: Patients aged more than 35
Exclusion: Patient with missing smoking/drinking status

All patients



min_date

min_date

ca_date

Lung/HNC

opd_y59

opd_y60

opd_y61

opd_y62

opd_y63

opd_y64

opd_y65

opd_y66

opd_y67

ipd

max_date

max_date

Last date of service

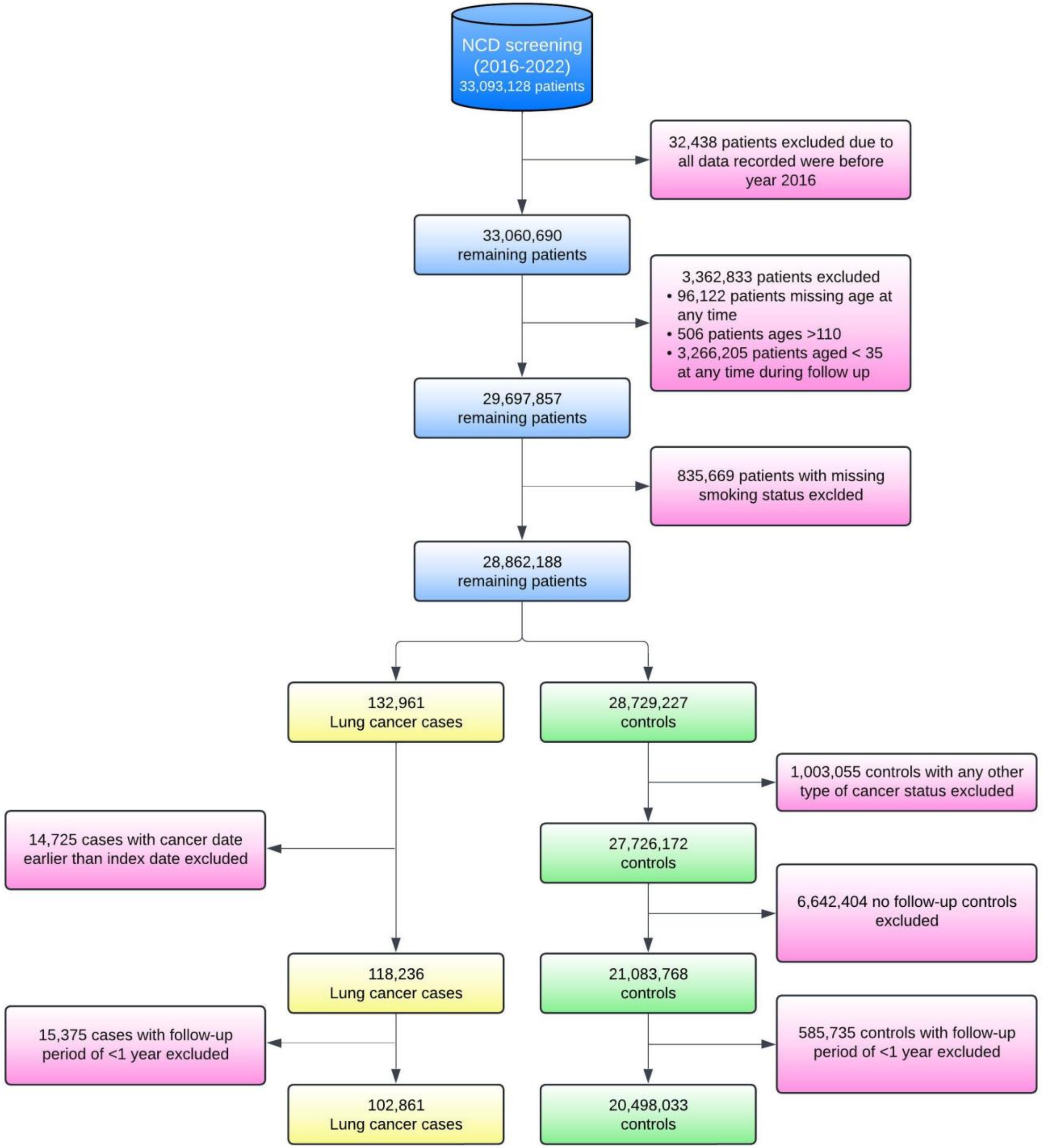
all_patient

Retrieve age missing from NCD screening

Create ID list of patient having any types of cancer

Exclusion: Controls having any types of cancer
C00-C97 Malignant neoplasms

Data flow of
lung cancer cohort



NCD screening (2016-2022)
33,093,128 patients

32,438 patients excluded due to all data recorded were before year 2016

33,060,690 remaining patients

3,362,833 patients excluded
- 96,122 patients missing age at any time
- 506 patients ages >110
- 3,266,205 patients aged < 35 at any time during follow up

29,697,857 remaining patients

835,669 patients with missing smoking status exclded

28,862,188 remaining patients

132,961 Lung cancer cases

28,729,227 controls

1,003,055 controls with any other type of cancer status excluded

14,725 cases with cancer date earlier than index date excluded

27,726,172 controls

6,642,404 no follow-up controls excluded

118,236 Lung cancer cases

21,083,768 controls

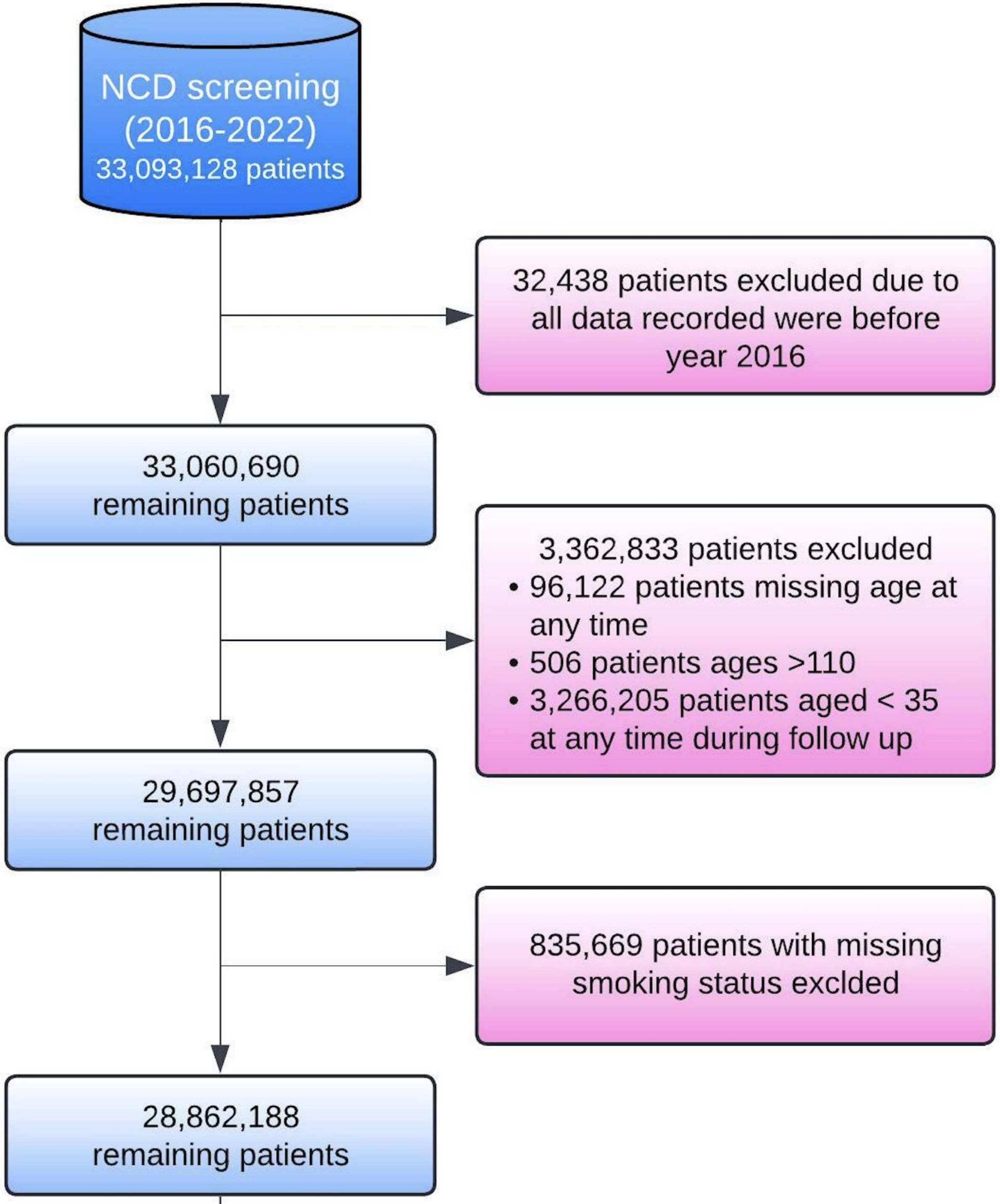
15,375 cases with follow-up period of <1 year excluded

585,735 controls with follow-up period of <1 year excluded

102,861 Lung cancer cases

20,498,033 controls

Data flow of lung cancer cohort



NCD screening (2016-2022)
33,093,128 patients

32,438 patients excluded due to all data recorded were before year 2016

33,060,690 remaining patients

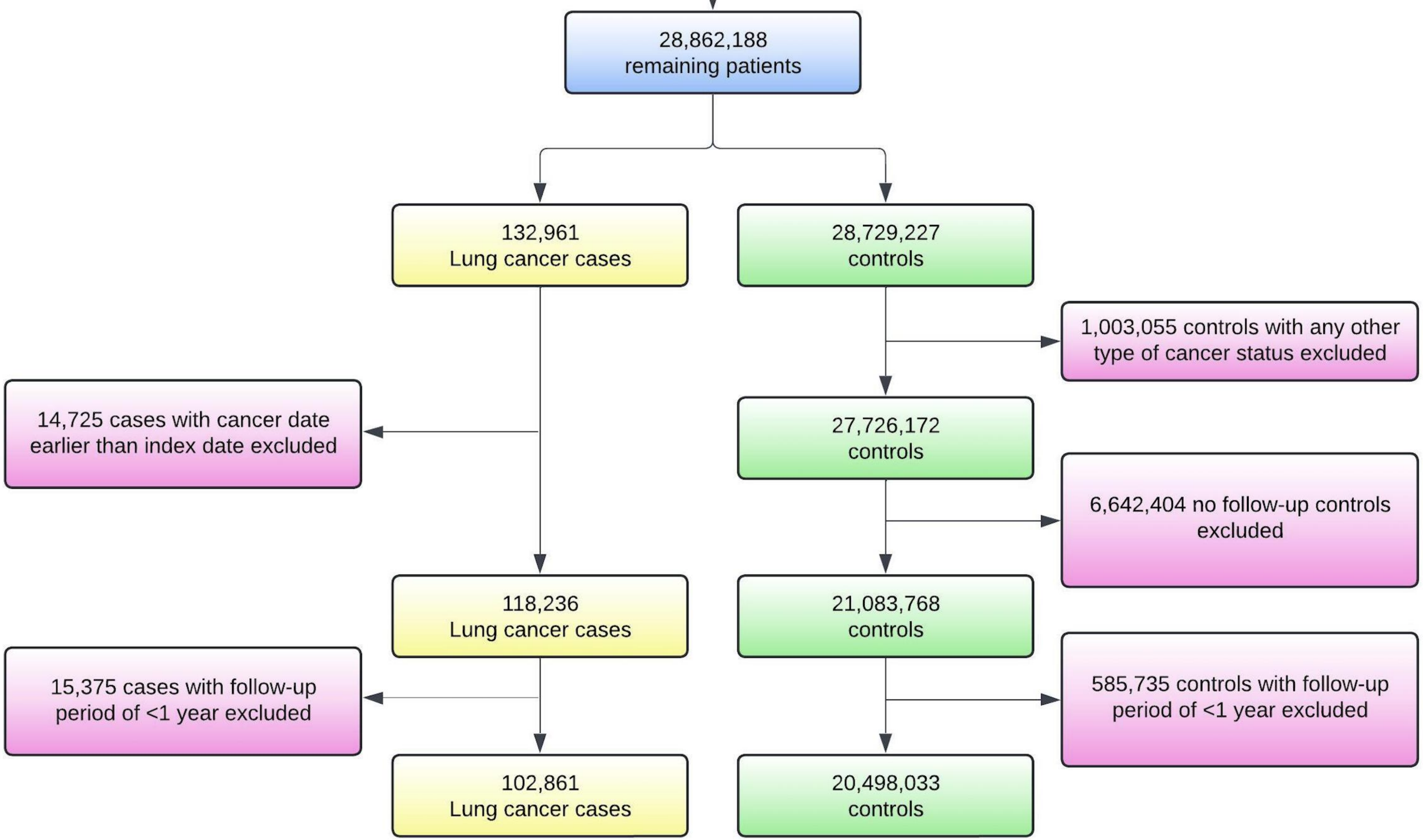
3,362,833 patients excluded
- 96,122 patients missing age at any time
- 506 patients ages >110
- 3,266,205 patients aged < 35 at any time during follow up

29,697,857 remaining patients

835,669 patients with missing smoking status exclded

28,862,188 remaining patients

Data flow of
lung cancer cohort



28,862,188
remaining patients

132,961
Lung cancer cases

28,729,227
controls

1,003,055 controls with any other type of cancer status excluded

14,725 cases with cancer date earlier than index date excluded

27,726,172
controls

6,642,404 no follow-up controls excluded

118,236
Lung cancer cases

21,083,768
controls

15,375 cases with follow-up period of <1 year excluded

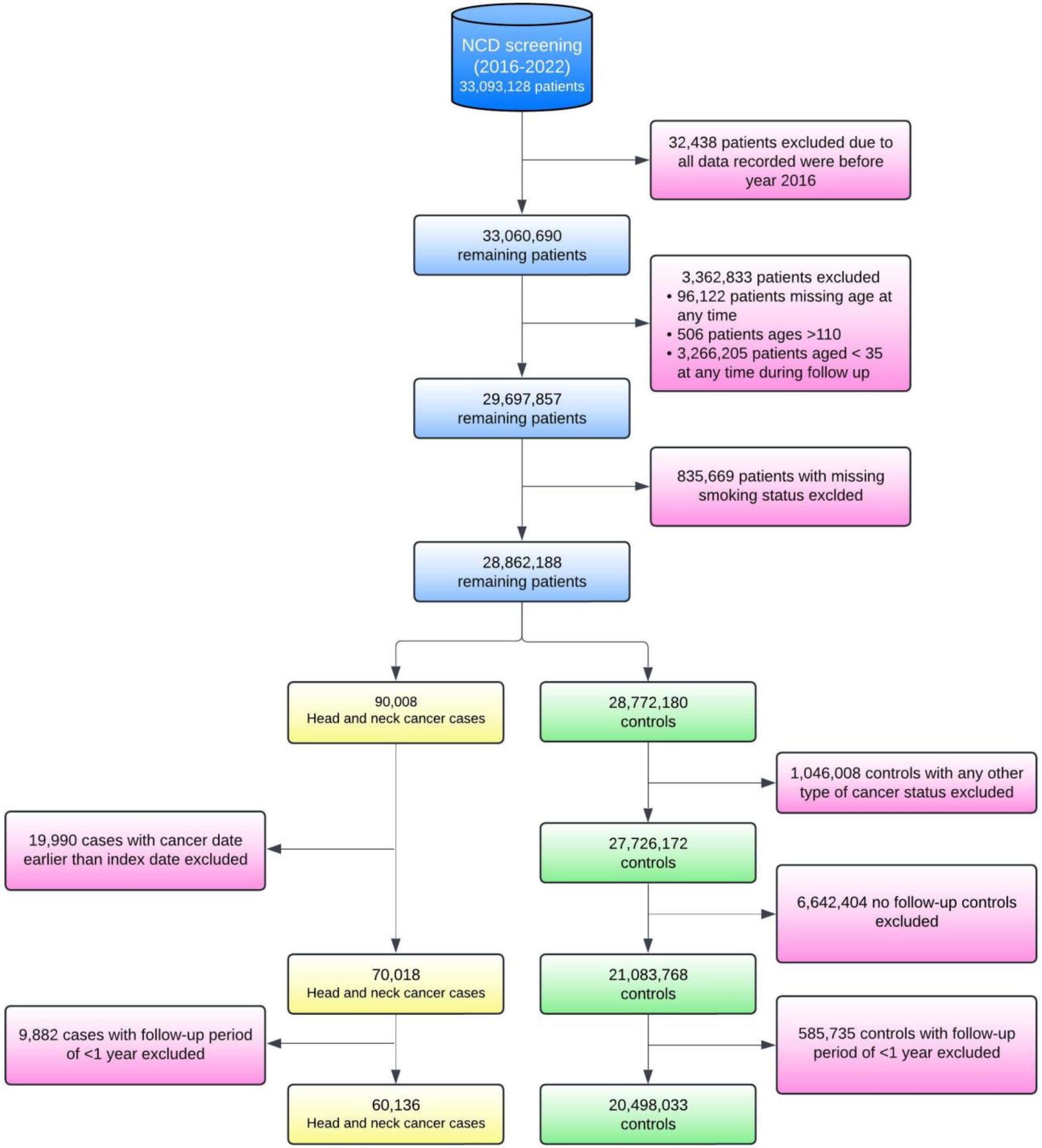
585,735 controls with follow-up period of <1 year excluded

102,861
Lung cancer cases

20,498,033
controls

Matched case-control
1:10 case control ratio
matching criteria - ±30 days follow-up period

Data flow of Head and neck cancer cohort



NCD screening (2016-2022)
33,093,128 patients

32,438 patients excluded due to all data recorded were before year 2016

33,060,690 remaining patients

3,362,833 patients excluded
- 96,122 patients missing age at any time
- 506 patients ages >110
- 3,266,205 patients aged < 35 at any time during follow up

29,697,857 remaining patients

835,669 patients with missing smoking status exclded

28,862,188 remaining patients

90,008 Head and neck cancer cases

28,772,180 controls

1,046,008 controls with any other type of cancer status excluded

19,990 cases with cancer date earlier than index date excluded

27,726,172 controls

6,642,404 no follow-up controls excluded

70,018 Head and neck cancer cases

21,083,768 controls

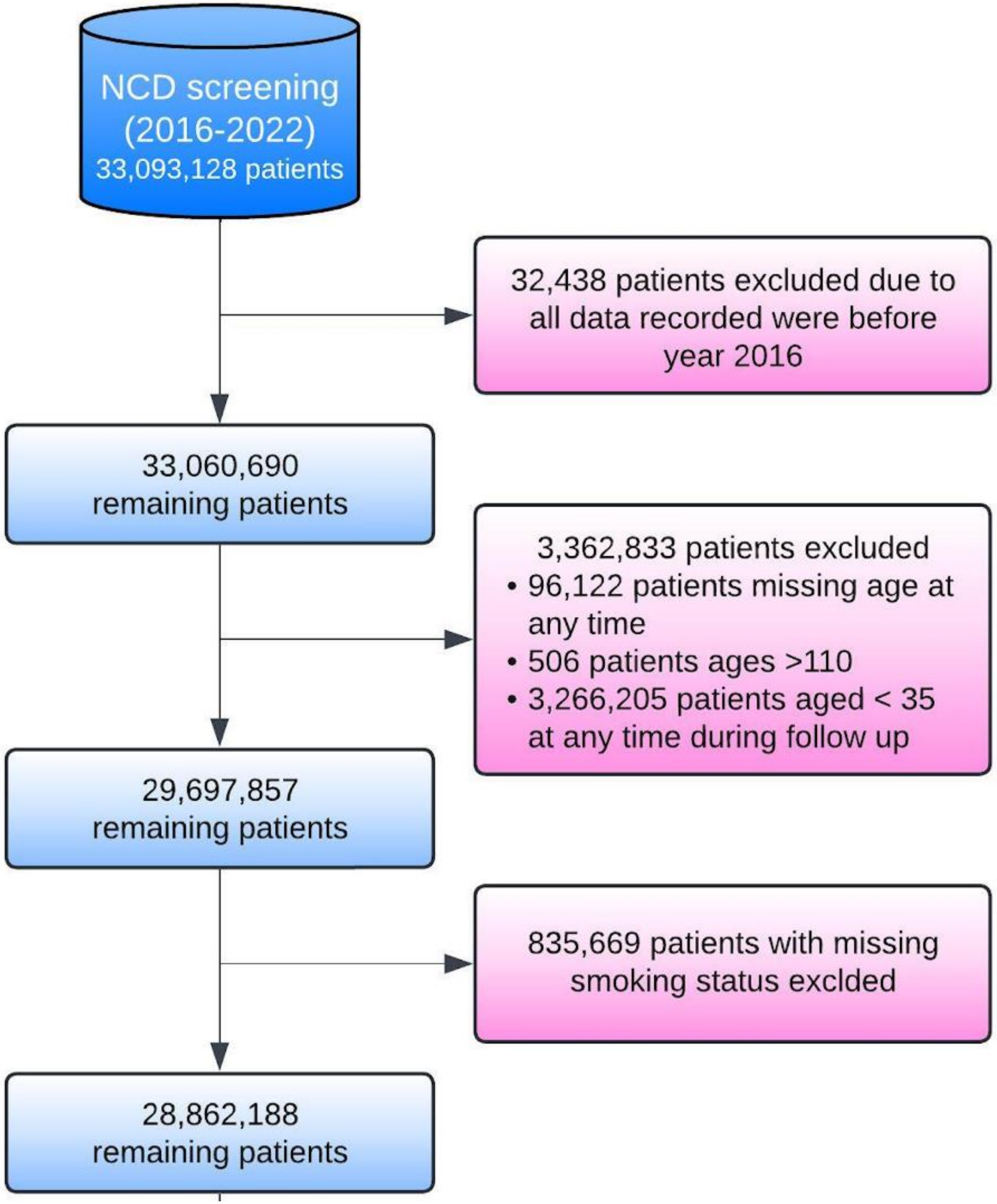
9,882 cases with follow-up period of <1 year excluded

585,735 controls with follow-up period of <1 year excluded

60,136 Head and neck cancer cases

20,498,033 controls

Data flow of Head and neck cancer cohort



NCD screening (2016-2022)
33,093,128 patients

32,438 patients excluded due to all data recorded were before year 2016

33,060,690 remaining patients

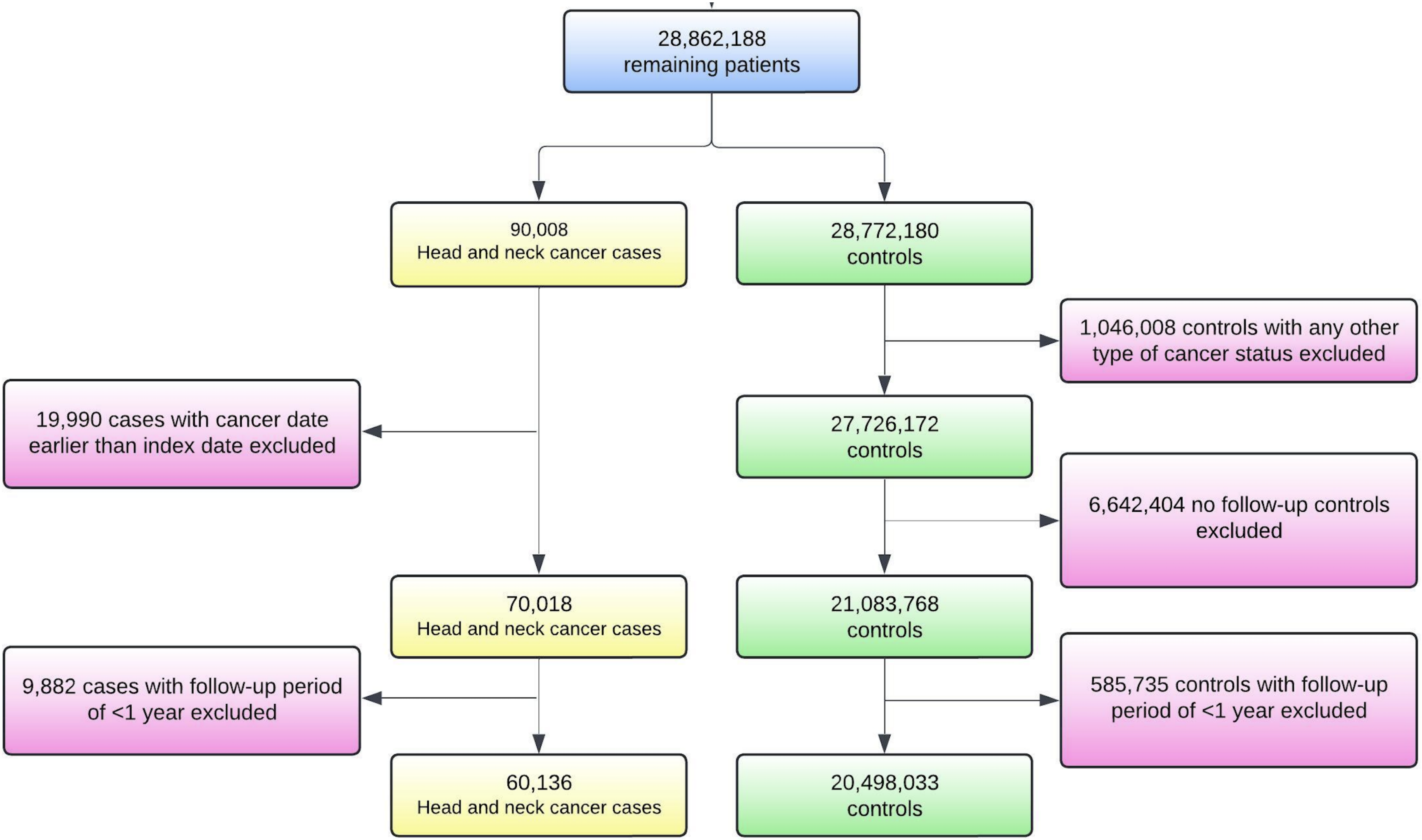
3,362,833 patients excluded
- 96,122 patients missing age at any time
- 506 patients ages >110
- 3,266,205 patients aged < 35 at any time during follow up

29,697,857 remaining patients

835,669 patients with missing smoking status exclded

28,862,188 remaining patients

Data flow of
Head and neck cancer
cohort



28,862,188
remaining patients

90,008
Head and neck cancer cases

28,772,180
controls

1,046,008 controls with any other type of cancer status excluded

19,990 cases with cancer date earlier than index date excluded

27,726,172
controls

6,642,404 no follow-up controls excluded

70,018
Head and neck cancer cases

21,083,768
controls

9,882 cases with follow-up period of <1 year excluded

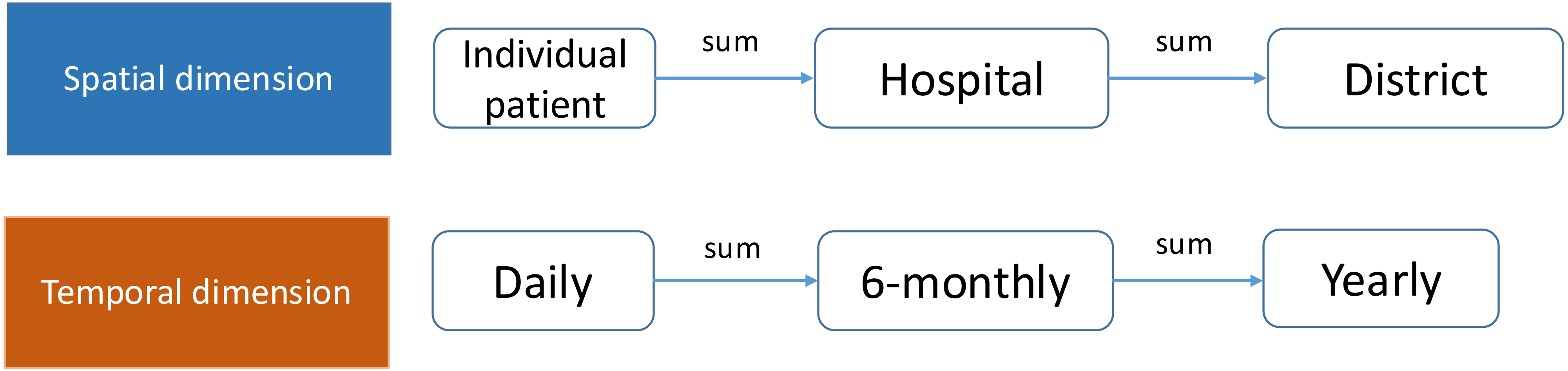
585,735 controls with follow-up period of <1 year excluded

60,136
Head and neck cancer cases

20,498,033
controls

Matched case-control
1:10 case control ratio
matching criteria - ±30 days follow-up period

Data preprocessing - Health outcome data


Spatial dimension

Individual patient → sum → Hospital → sum → District

Temporal dimension

Daily → sum → 6-monthly → sum → Yearly

- For the aggregate level analysis, district-level annual incidence rate of lung, head, and neck cancers will be calculated.
  - $IR = \frac{Number\ of\ new\ cases}{Total\ population} \times\ 100{,}000$

Data preprocessing - Health outcome data

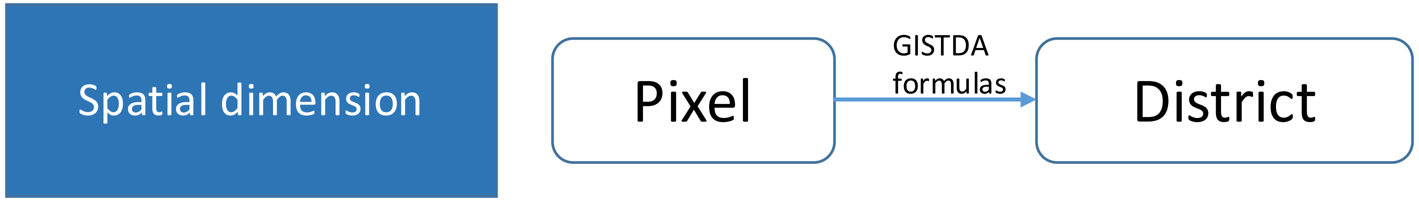
- For the individual level analysis, lumping method will be applied to lump multiple records belonging to the same individual, using time period of one year.
- The data linkage will be carried out between outcome and exposure data using the district-year codes.

Exposure data sources

- **PM2.5**
  - **Himawari** satellite - data for 10 years from 2015 to 2024.
    - Spatial resolution - approximately 6 x 6 km
    - Temporal dimension - hourly
  - **MODIS** satellite, data 21 years from 2002 to 2024.
    - Spatial resolution - 10 x 10 km
    - Temporal dimension – daily
- **Other** exposure data - NASA's dataset: Nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3)

Preprocessing plan for PM2.5 data

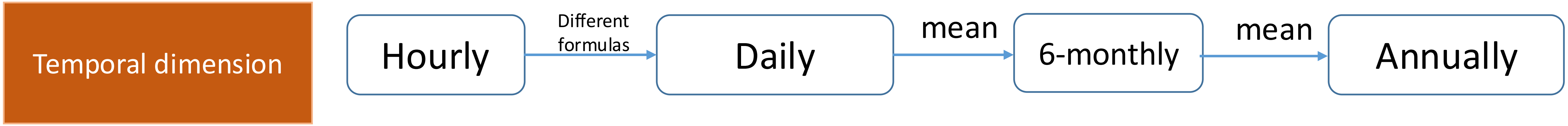
GISTDA has already done calculation of district-level PM2.5 concentrations using the pixel-level data.



Spatial dimension

Pixel → GISTDA formulas → District

Temporal aggregation to be applied in this study.

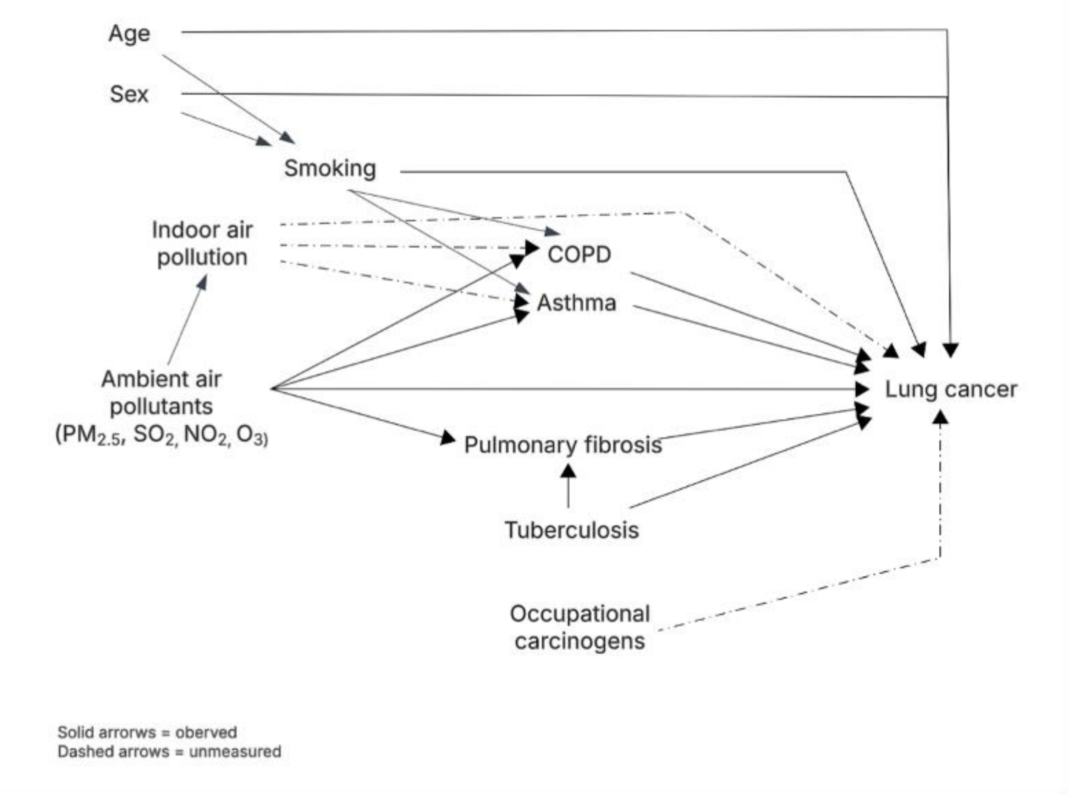


Temporal dimension

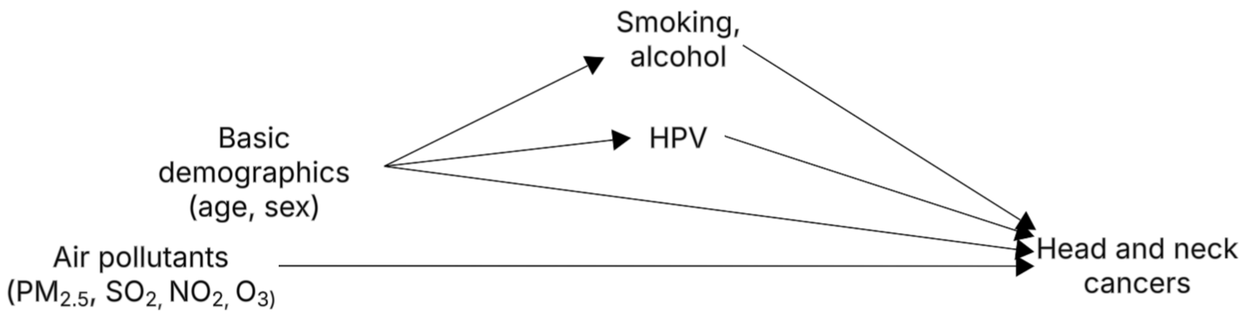
Hourly → Different formulas → Daily → mean → 6-monthly → mean → Annually

Data analysis plan (cont.)

- Individual level analysis – statistical models
  - Weighted logistic regression model
  - Hierarchical mixed-effects logistic regression model
  - ML models:
    - Random forest
    - XGBoost
    - Bayesian Additive Regression Trees (BART)



Causal diagram for the supposed effect of PM2.5 and other pollutants on lung cancer


Smoking, alcohol
Basic demographics (age, sex)
HPV
Air pollutants ($PM_{2.5}$, SO_2, NO_2, O_3)
Head and neck cancers

Causal diagram for the supposed effect of PM2.5 and other pollutants on head and neck cancers

Double/Debiased ML (DML)

- **Objective:** to estimate the **total causal effect** of long-term PM2.5 exposure on lung cancer and HNC.

- **Approach:**
  - DML with Orthogonal Random Forest (ORF)
  - Compatible with **continuous** PM2.5 exposure and cancer **binary** outcome
  - **Random forest** and ridge/lasso debiasing
  - Orthogonalization + cross-fitting for valid inference
- **Outputs:**
  - Average causal effect
  - Conditional causal effects for sub-groups (age, sex, smoking status)

Thank You