2020-2025 JSPS A3 Foresight Program (Grant No. JPJSA3F20200001)

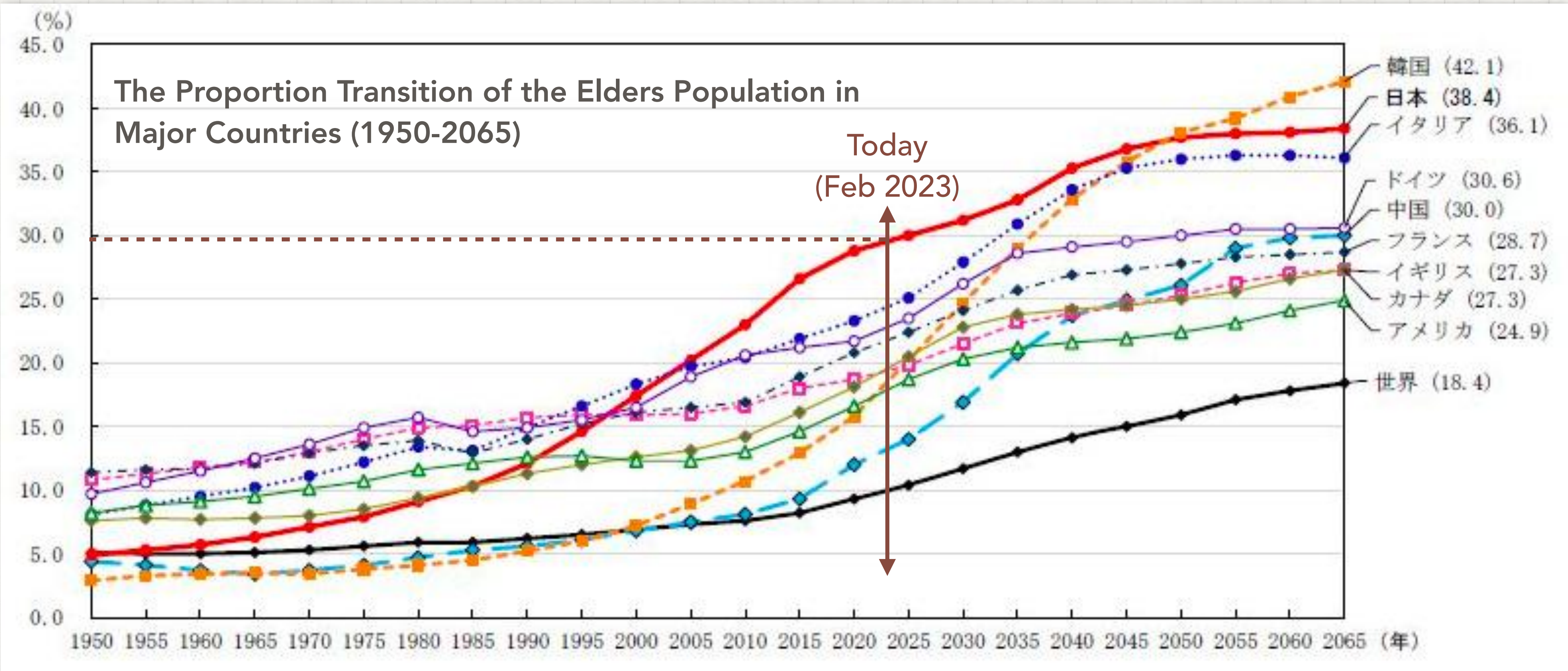
Health Data Analysis by Causal Discovery for Supporting Policy Formation and Municipal Problem-Solving

Feb 15, 2023

Ou Deng, Shoji Nishimura, Atsushi Ogihara, Qun Jin
Graduate School of Human Sciences, Waseda University
dengou@toki.waseda.jp

Background

The unstoppable progress of aging.
Today, Japan faces the most serious situation in the world.



On 2065

Korea 42.1%
Japan 38.4%
Italy 36.1%

Germany
China
France
United Kingdom
Canada
United States

World

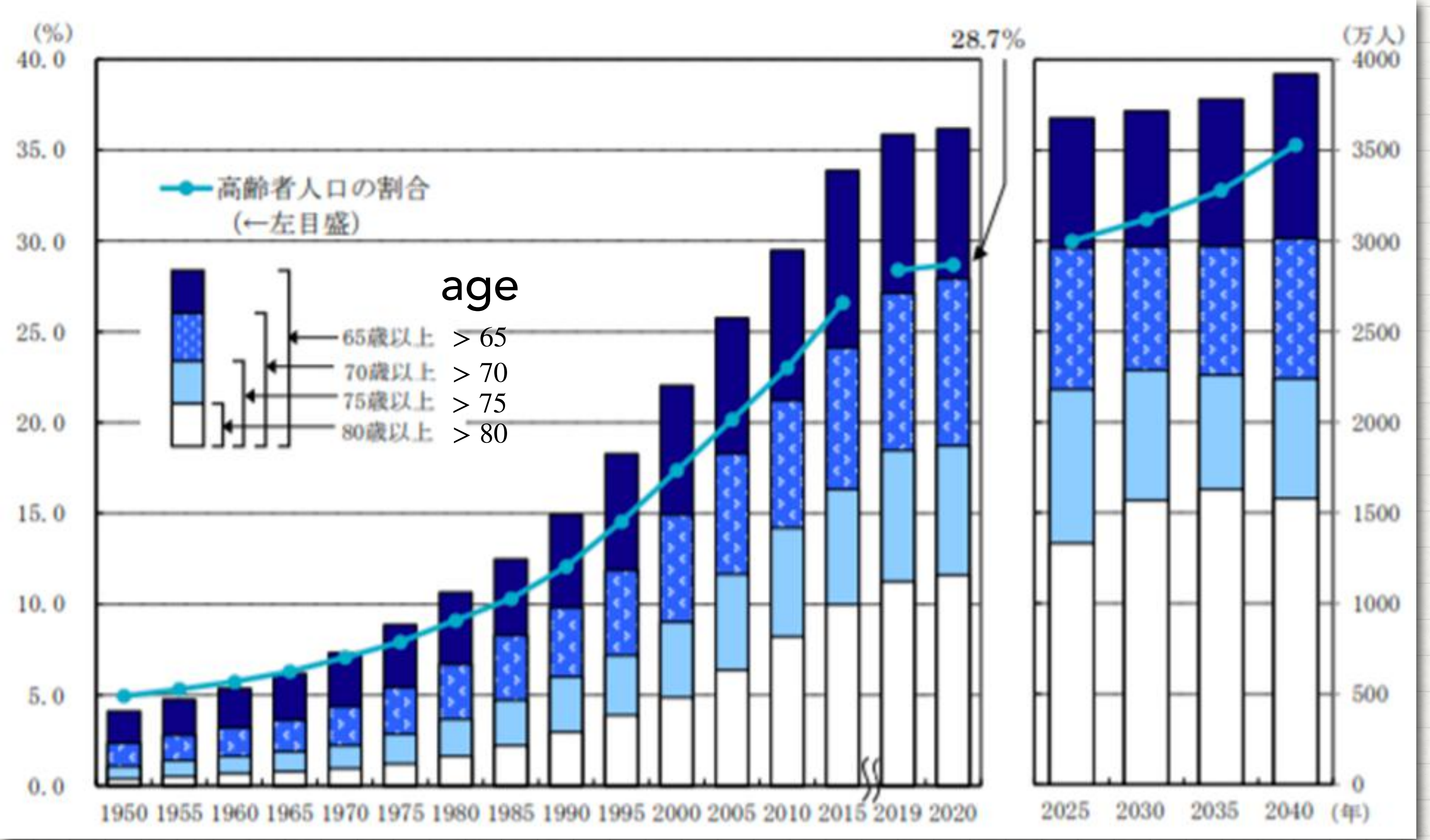
Resource: World Population Prospects: The 2019 Revision (United Nations)

United Nations

Note: Elderly = people over 65-year-old (Japan)

Background

Elders Population Structure and the Proportion Transition of Japan (1950-2040)



Unit: 10k people

Japan Now:

>65 yr-old 36.2 million (29%)
>75 yr-old 28.0 million (22%)
>85 yr-old 6.2 million (5%)
>90 yr-old 2.4 million (2%)

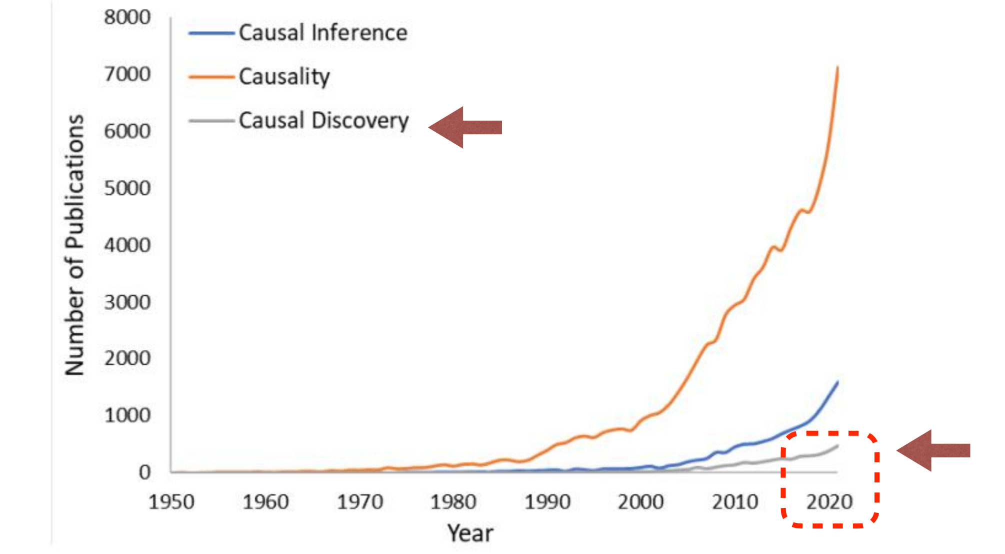
(): the percentage of the population (≈120 million)

The elders population (>65yr-old) compares to the labor population(15-64 yr-old) is 13% in Japan.

Resource: Ministry of Internal Affairs and Communications of Japan, 2020

Why Need Causal Discovery in Healthcare Research?

- Too expensive or difficult, or ethically impossible to perform certain controlled experiments.
- To identify the underlying causal relationships between variables, such as the relationship between a particular risk factor and a disease outcome.
- Help to identify new and unexpected relationships between variables that may have gone unnoticed in traditional research methods.



Yearly publications for causal inference and causality (data derived from Scopus)

A. Rawal et al., Causality and Machine Learning Review. DTIC. Nov 2022.

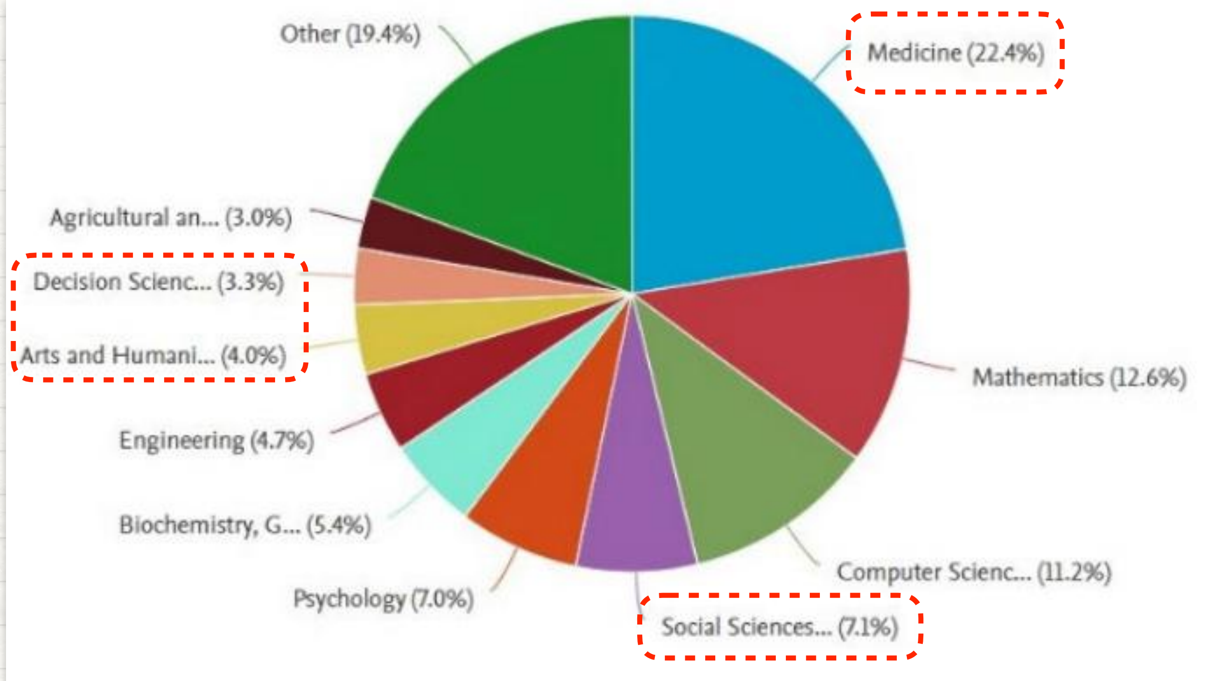


Fig. 4 Applications for causality (data derived from Scopus)

Related Causal Discovery Work for Japan Healthcare Data

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Causal relations of health indices inferred statistically using the DirectLiNGAM algorithm from big data of Osaka prefecture health checkups

Jun'ichi Kotoku, Asuka Oyama, Kanako Kitazumi, Hiroshi Toki, Akihiro Haga, Ryohei Yamamoto, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, Keiichi Yamamoto, Toshiki Moriyama

Published: December 23, 2020 • https://doi.org/10.1371/journal.pone.0243229

Data:

- Health checkup data (anonymized) of Osaka prefecture during 2012-2017
- Target 60s/70s/80s age groups with 30k+ samples each

Age	Men	Woman
50-59	20,316	26,654
60-69	69,892	109,529
70-79	97,327	131,036
80-89	32,594	46,906

Method:

- Direct LiNGAM [1]
  (Linear Non-Gaussian Acyclic Model)

Target health concerning variables:

- Selected 11 indicators [2] for elders
  (details on next page)

[1] Shimizu, Y., et al. A linear non-Gaussian acyclic model for causal discovery. Journal of machine learning research, 7:2003-2030. 2006

[2] J. Kotoku et al. Causal relations of health indices inferred statistically using the DirectLiNGAM algorithm from big data of Osaka prefecture health checkups. PLOS ONE, 15(12), 2020

Most Critical Diseases in the Elderly Population

The basis for selecting the aforementioned 11 indicators for causal discovery is grounded in prior knowledge, specifically from investigations into elderly populations residing in Osaka.

Heart disease and stroke
심장병과 뇌졸중
心臓病と脳卒中

Liver diseases
간질환
肝疾患

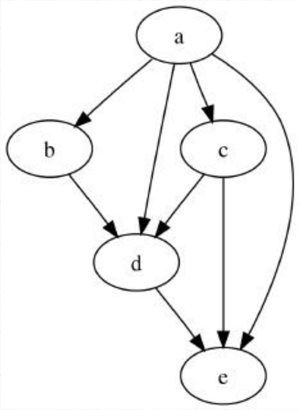
/ˌdaɪ.əˈbiː.t̬əs/
Diabetes
당뇨
糖尿病

/nɪˈfrɑːp.ə.θi/
Diabetic nephropathy (DN)
당뇨성 신염
糖尿病性腎症

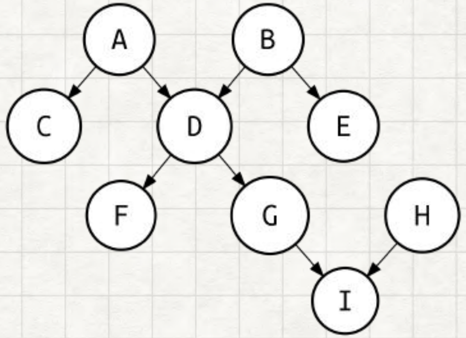
11 selected indicators	Chinese	Korean	What's it?
Systolic blood pressure (**sBP**)	收缩压	시스토릭 혈압	The **highest pressure** reaches the arteries when the heart beats and pumps blood. Normal: <120mmHg and Elevated: >140mmHg.
Low-density lipoprotein cholesterol (**LDL**)	低密度脂蛋白胆固醇	저밀도 지질 콜레스테롤	A type of blood fat called "**bad cholesterol**" deposits on blood vessels' walls to increase the risk of cardiovascular disease.
High-density lipoprotein cholesterol (**HDL**)	高密度脂蛋白胆固醇	고밀도 지질 콜레스테롤	A type of blood fat called "**good cholesterol**" removes other fats, particularly the LDL.
Triglyceride (**TG**)	甘油三酯	트리글리세라이드	A type of blood fat **to store excess energy**. High-level TG increases the risk of heart disease and stroke.
Glutamic Oxaloacetic Transaminase (**GOT**)	谷氨酸草酸转氨酶	글루타믹 옥사로아세틱 트랜사민에이즈	Blood biochemical marker; an important **indicator of liver health**. Elevated GOT levels indicate liver disease or other organ damage.
Gamma-glutamyl transpeptidase (**γGT**)	γ-谷氨酰转换酶	감마 글루타믹 트랜스펩타다이즈	Another important **indicator of liver health**, also concerning with cardiovascular disease, type 2 diabetes, and certain types of cancer.
Glutamic Pyruvic Transaminase (**GPT**)	谷氨酸丙酮转氨酶	글루타믹 프루비카 트랜사민에이즈	Another important **indicator of liver health**. Elevated GPT levels indicate liver disease or other organ damage.
Body Mass Index (**BMI**)	体质指数	체질 지수	An indicator of **fatness** to evaluate the risk of heart disease, stroke, and diabetes. Normal BMI in the range [18.5, 24.9].
Fasting blood glucose level (**fBG**)	空腹血糖水平	식전 혈당 수치	**Glucose (sugar) in blood** after fastening(typically 8 hr) to diagnose diabetes and pre-diabetes.
Hemoglobin A1c (**HbA1c**)	血红蛋白A1c	헤모글로빈 A1c	**2 to 3 months average blood glucose level** to diagnose and monitor diabetes and assess a person's glucose control.
Height	身高	신장	Serve other indicators in analysis.

Causal Discovery: Fundamental Concepts

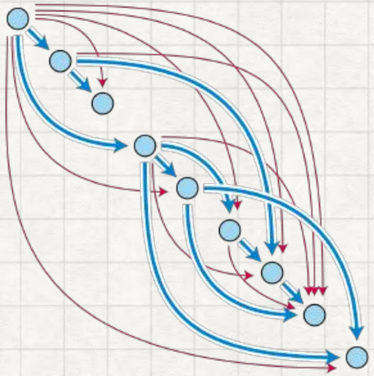
Directed Acyclic Graph (DAG)



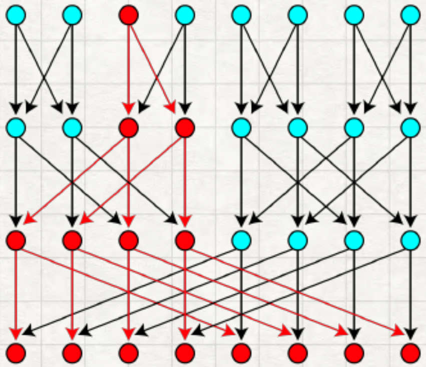
a. Natural DAG



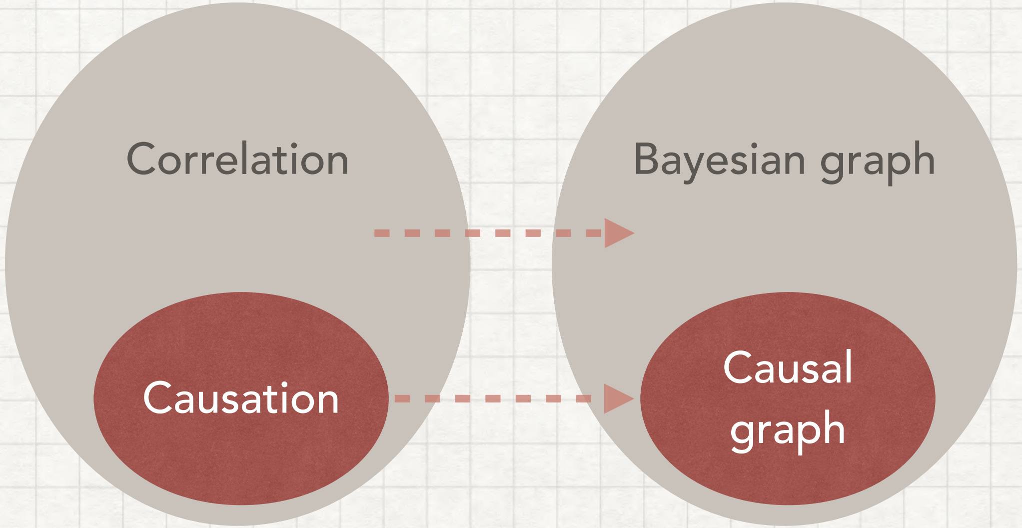
b. Polytree



c. Transitive closure (blue)



d. Multitree

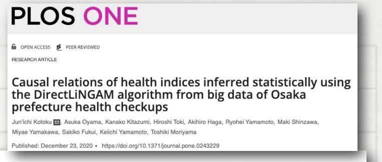


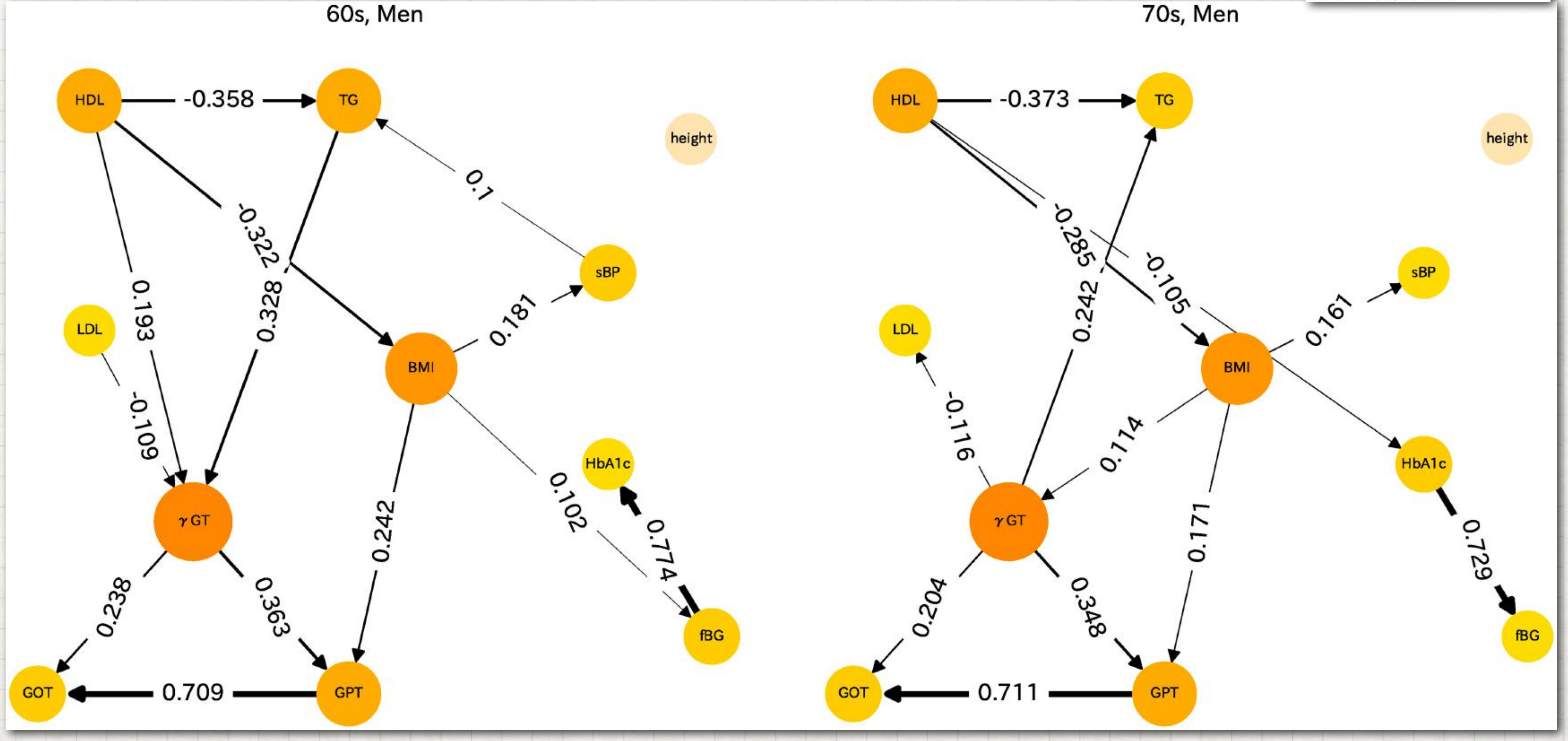
Relationship of correlation and causation

- A causal graph is the most concise bayesian graph, i.e., the most straightforward one.
- A causal graph is always a Directed Acyclic Graph(DAG) as the research object.

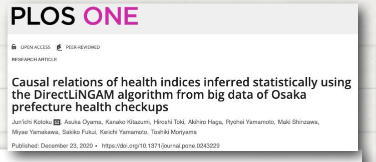
K. Thulasiraman et al., Graphs: Theory and Algorithms, John Wiley and Son, p.118. 1992
https://en.wikipedia.org/wiki/Directed_acyclic_graph

Causal Discovery Results: Men of 60s and 70s


PLOS ONE
OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE
Causal relations of health indices inferred statistically using the DirectLiNGAM algorithm from big data of Osaka prefecture health checkups
Jun'ichi Kotoku, Asuka Oyama, Kanako Kitazumi, Hiroshi Toki, Akihiro Haga, Ryohei Yamamoto, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, Keiichi Yamamoto, Toshiki Moriyama
Published: December 23, 2020 • https://doi.org/10.1371/journal.pone.0243229



Causal Discovery Results: Women of 60s and 80s


PLOS ONE

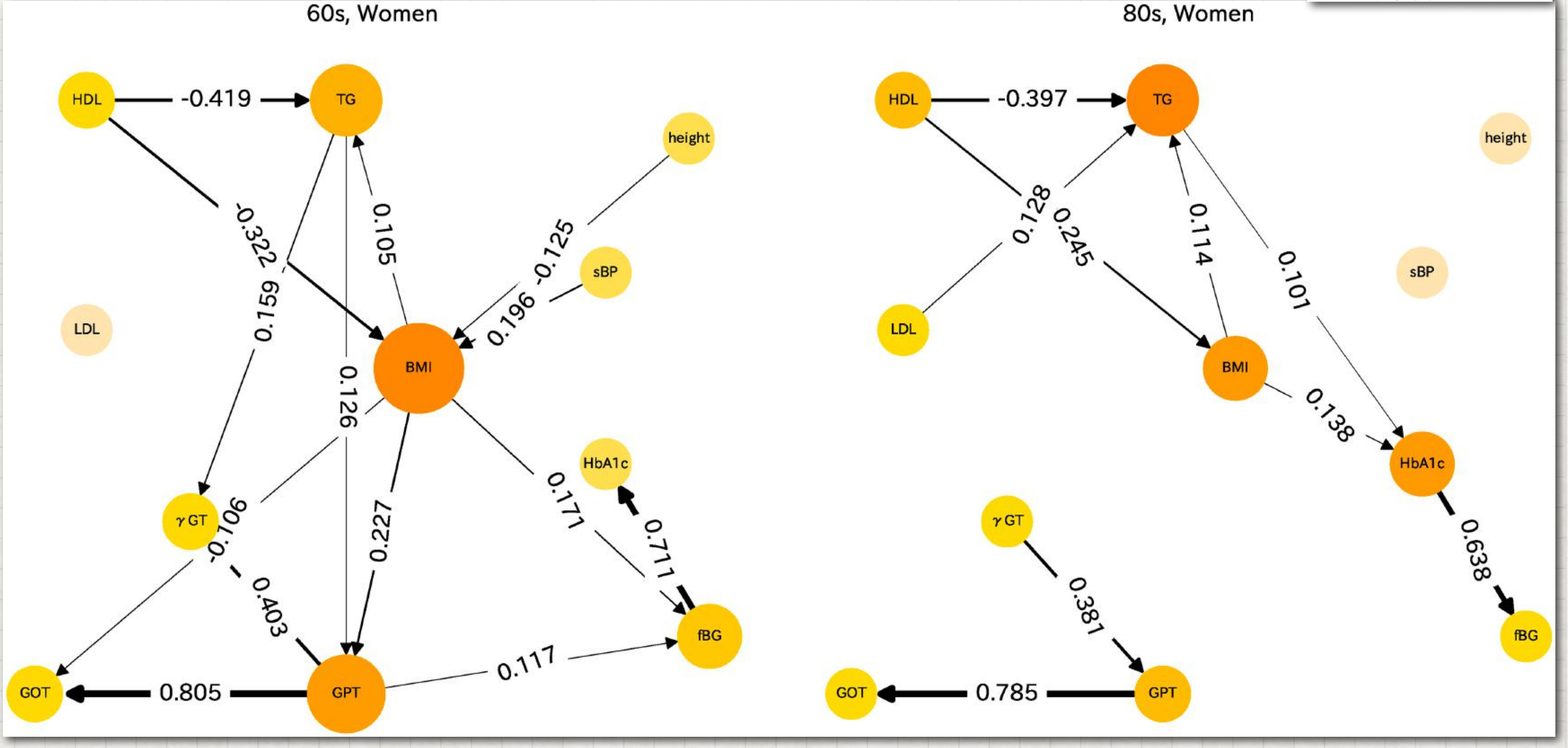
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

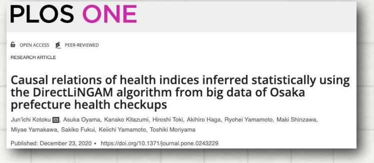
Causal relations of health indices inferred statistically using the DirectLiNGAM algorithm from big data of Osaka prefecture health checkups

Jun'ichi Kotoku, Asuka Oyama, Kanako Kitazumi, Hiroshi Toki, Akihiro Haga, Ryohei Yamamoto, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, Keiichi Yamamoto, Toshiki Moriyama

Published: December 23, 2020 • https://doi.org/10.1371/journal.pone.0243229



Conclusion with Causal Relationship fundings


PLOS ONE

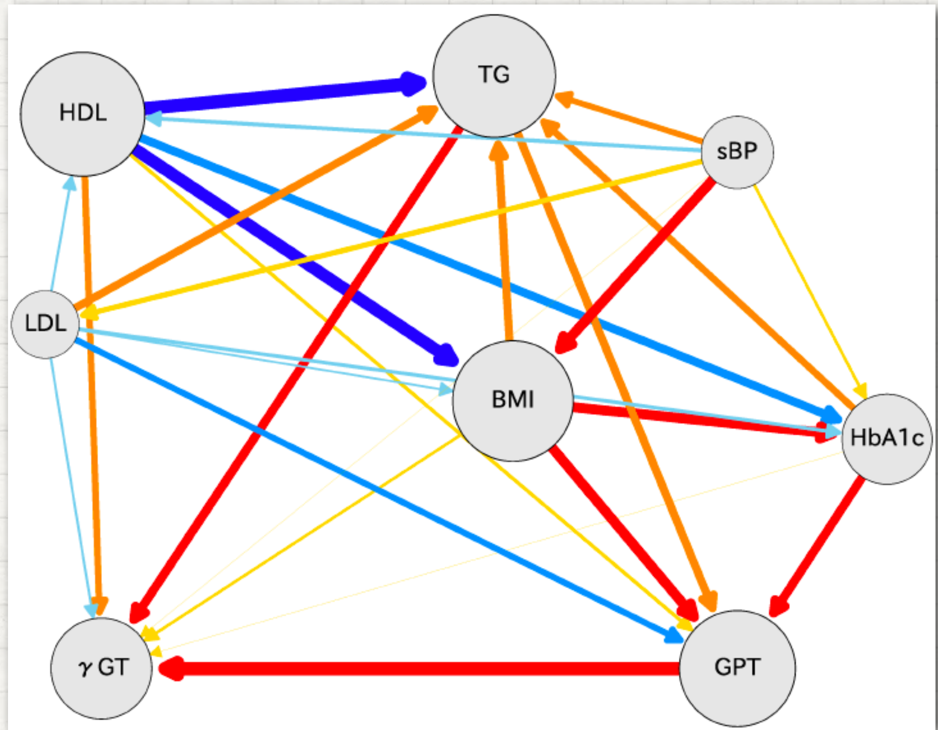
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Causal relations of health indices inferred statistically using the DirectLiNGAM algorithm from big data of Osaka prefecture health checkups

Jun'ichi Kotoku, Asuka Oyama, Kanako Kitazumi, Hiroshi Toki, Akihiro Haga, Ryohei Yamamoto, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, Keiichi Yamamoto, Toshiki Moriyama

Published: December 23, 2020 • https://doi.org/10.1371/journal.pone.0243229



Causal graph sample of age 70s women,
by Direct LiNGAM

For samples of 20,000 or smaller, errors in causality become large. The causality relations become fragile.

The role of HDL on BMI and TG (the dark blue thick arrows in the figure) is quite important and true for all age groups, both men and women. However, the role of LDL on other indices is small.

Several interesting causal relations were found to be quite robust:

- HDL strongly influences BMI and TG; this relation is robust in all age groups.
- LDL is quite independent.
- sBP influences BMI.
- BMI influences fBG (HbA1c) and GPT (**γ**GP).
- TG influences GPT (**γ**GP).
- fBG and HbA1c are correlated strongly, but the causal order is fragile.
- GOP is influenced both by GPT and by **γ**GP.

Note: some fundings need assess in greater detail by specifically examining these indices using other statistical methods.

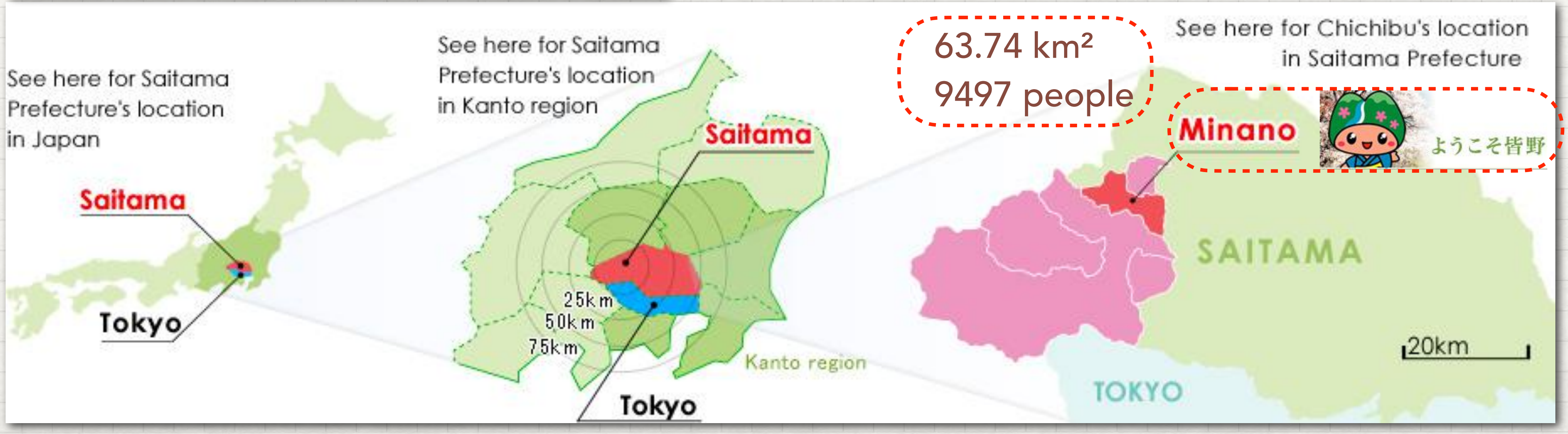
Next Work: Do Further beyond Causal Diagram

Review this research PJ title, which cooperates with local government.

(What) Position: Health Data Analysis
(How) Method: by Causal Discovery
(Why) Purpose: for Supporting Policy Formation and Municipal Problem-Solving

Research PJs comparison	Osaka	Minano
Scale	Metropolitan	Local town
Population	About 8,823,000	About 9,000
Data	Health checkup data only Selected groups and indicators	All health-related data All citizens
Causal discovery purpose	Causal graph	Causal graph, Tracing, Simulating etc
Project	Academic independed	Academic co-work with local government
Support policy formation	Possible, but week	Possible, might be strong
Municipal problem solving	Possible, but week	Possible, aiming to resonable effect
Extended social experiment	No	Yes
Duration	Half year (Finished in 2020)	Three years (Started in Nov, 2022)

Minano-machi, Japan



Resource: **Minano Tourism Association** https://www.minano.gr.jp/en/

Minano-machi, Japan

Resource: **Minano Tourism Association** https://www.minano.gr.jp/en/

Minano-machi, Japan

https://youtu.be/UruJDhMHAM0 YouTube

Resource: **Minano Tourism Association** https://www.minano.gr.jp/en/

Health Data Analysis by Causal Discovery for Supporting Policy Formation and Municipal Problem-Solving

Research Plan

(I) Data preparation phrase

Data sources

NHI(KDB)

MUNICIPAL HMS

Anonymization

Data storage

DB for analyzing

Data master
- Receipt(Treatment+Dispensing)
- Special health check-up
- NIH(KDB)
- Elderly care

Filtering

(II) Data analysis phrase

Problems Input
- Specific diseases
- Regional health issues
- Preventive measures

Mechanism of Causal Discovery
- Time series data
- Non-time series data

Data analyzing

Visualization of Analysis Results
&
On-Site Evidence Verification

(III) Outcomes evaluation phrase

(Improvement of Outcomes)

Results output
- Causal graph of health data
- Evidence for policy formation
- Solution to municipal problems

Note:
- NHI(KDB): Database system managed by National Health Insurance Organizations(NHI).
- Municipal HMS: Local Health Management System managed by the municipal government.

Phrase-I: Data preparation

General processes

NHI(KDB)

(National DB, managed by the central government)

Data source	National Health Insurance Organizations (NHI) (Medical, Diagnosis Proceduer Combination (DPC), Dental, and Dispensing)
Data duration	2016-2022 (7 years)
Record amount	**230,000** records (Medical only) **30,000** records (Health checkup, etc)
Main info	Insured information, Injury/Disease information, Medical information

MUNICIPAL HMS

(Local DB, managed by the local government)

Data source	Municipal Health Management System (HMS)
Data duration	2016-2022 (7 years)
Record amount	**6,500** records
Main info	Basic health checkup information, Osteoporosis checkup results, etc.

- All data is personal data, managed by individual social insurance numbers.
- A lot of the data is input manually by healthcare execution organizations with unavoidable errors.
- Many sub-databases can not organize data automatically.
- Other problems.

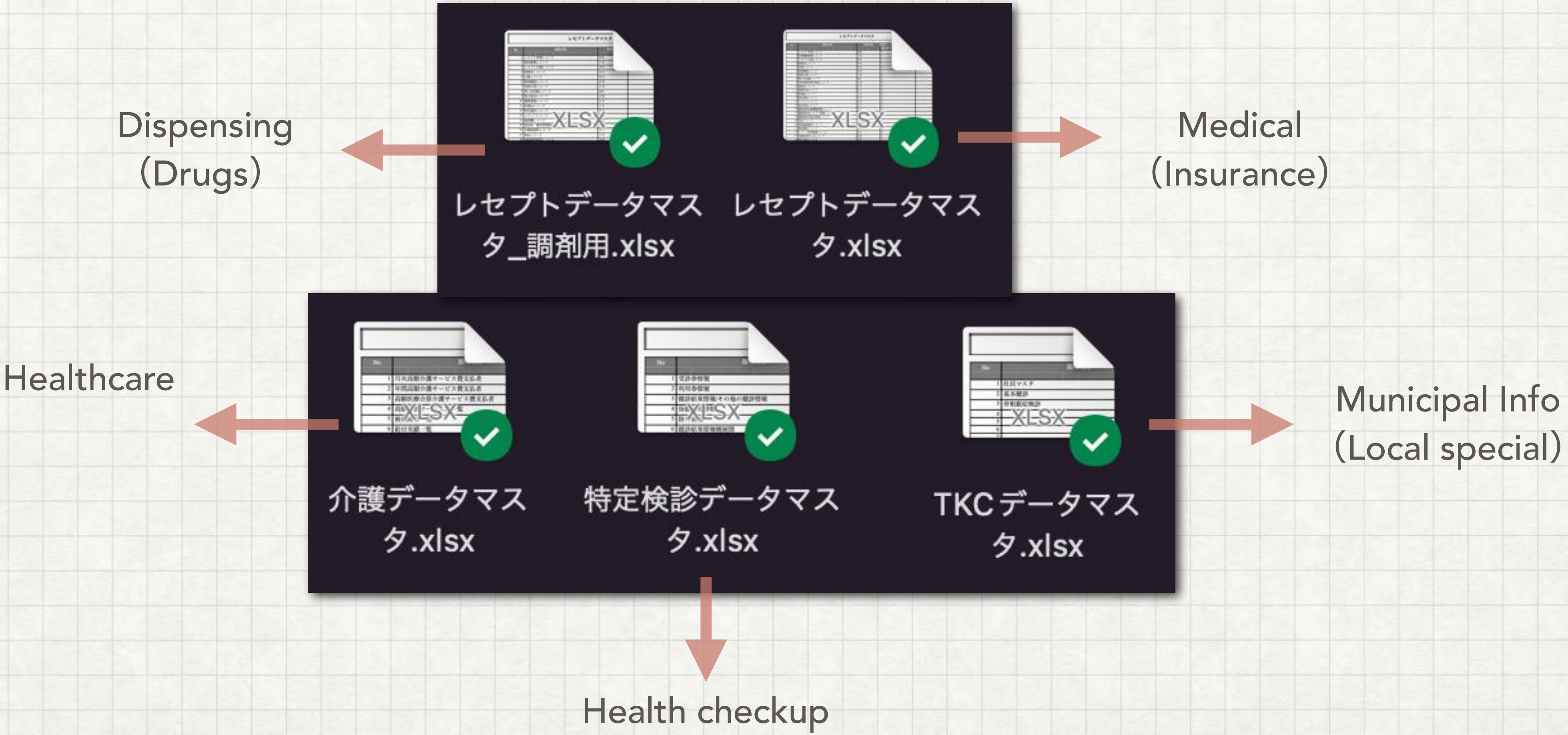
- Data washing
- Data master files
- Anonymized processes

Local DB for analyzing (build new), managed by the PJ research team.

Phrase-I: Data preparation

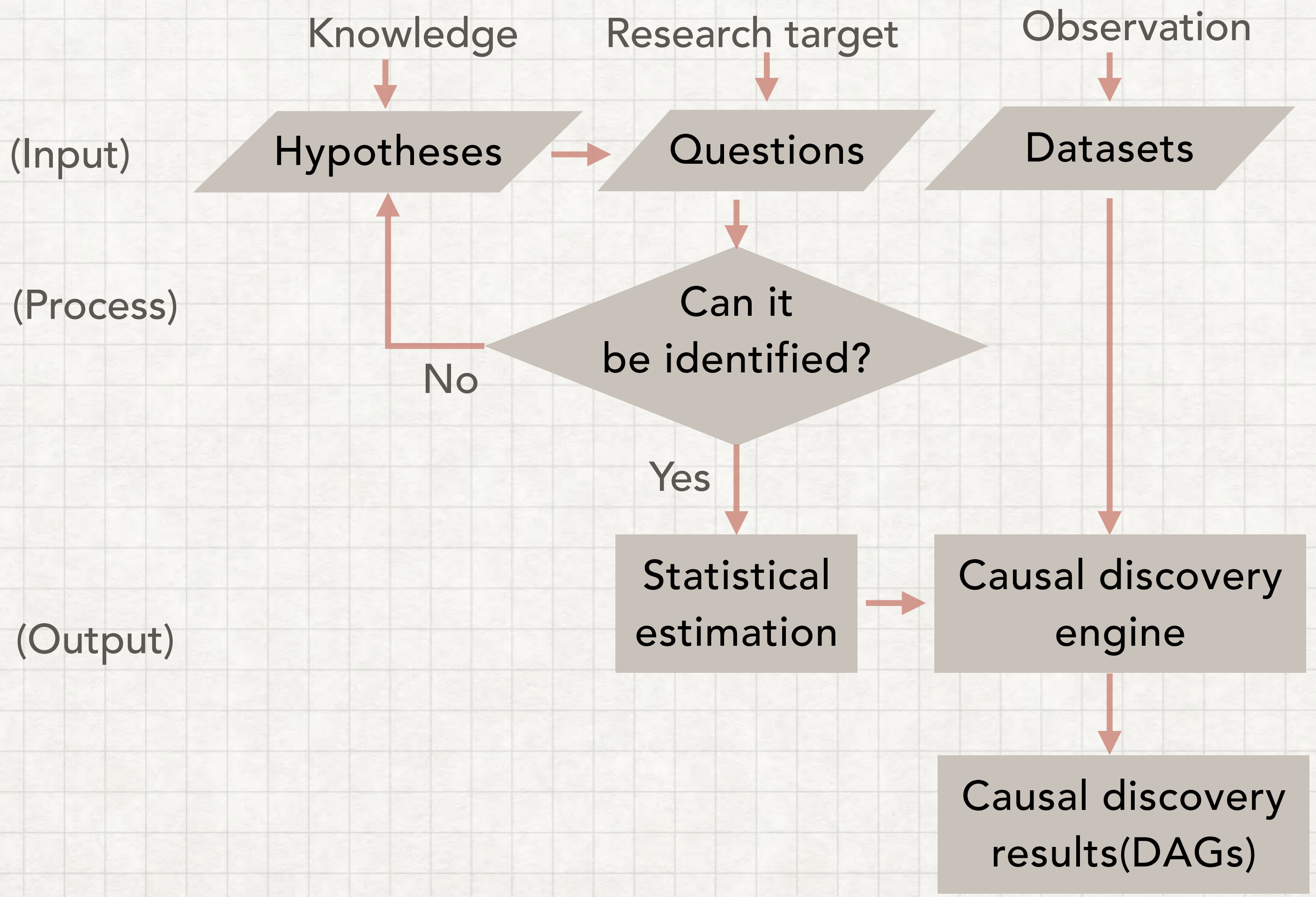
Data master files (Entity Relationship (ER), Indicators category)

* In processing.



Phrase-II: Data analysis

Causal discovery routine



Local Problems: requests from the elder healthcare frontline

Key risks of elders	Chinese	Korean	Japanese
Dementia	痴呆症	치매	認知症
Osteoporosis	骨质疏松症	골절마비증	骨粗しょう症
Diabetes	糖尿病	당뇨	糖尿病
Diabetic nephropathy	糖尿病肾病	당뇨성 신염	糖尿病性腎症
Cardiovascular and cerebrovascular	心脑血管疾病	심혈관질환, 뇌혈관질환	心脳血管疾患

The above diseases concern:

- Do people with ****** also suffer from some specific diseases?
- How do lifestyle and family situations affect ******?
- What's the effective early intervention for ****** prevention?

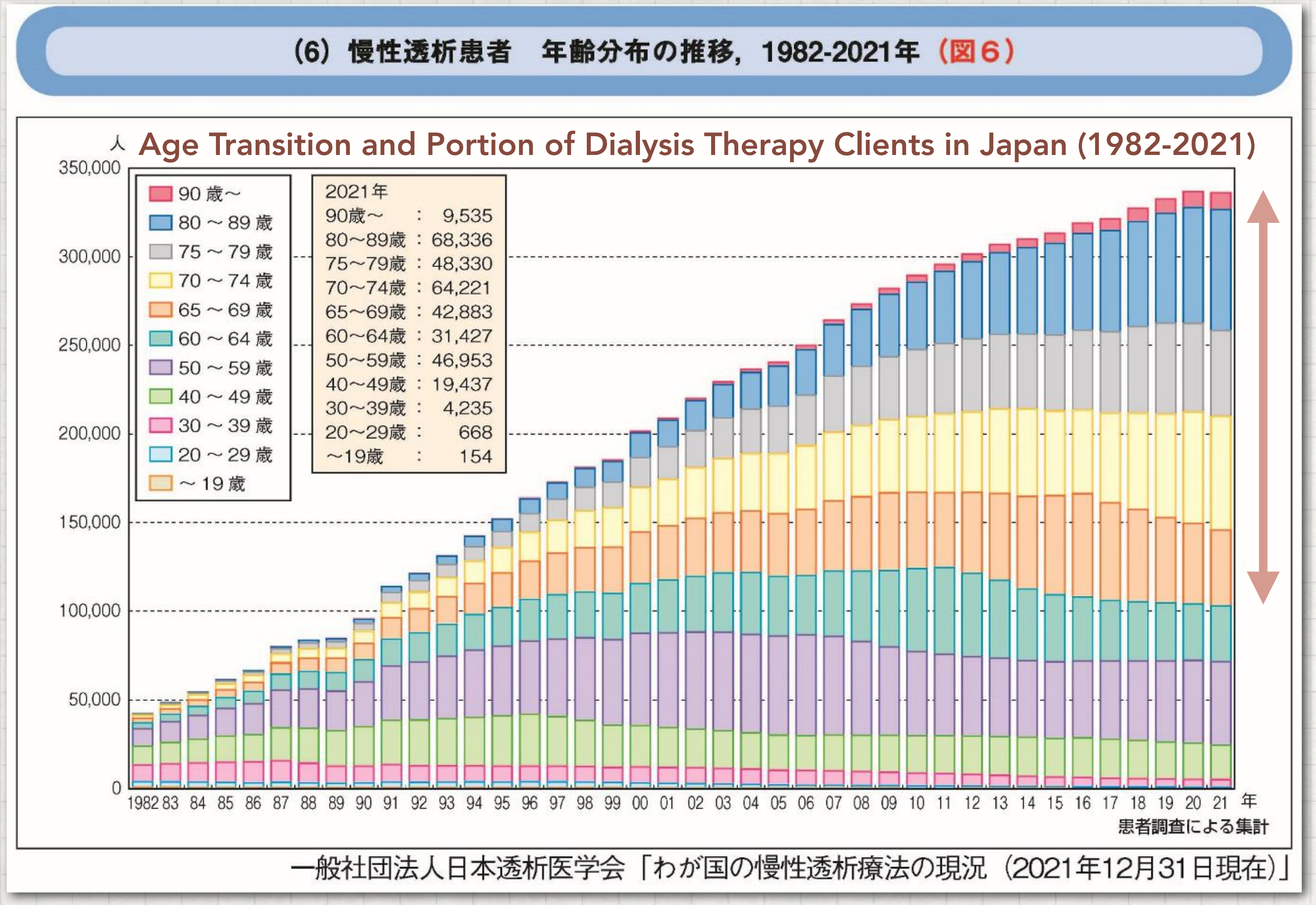
  etc.

The frontline healthcare financial control concerns:

- How to deduce the necessity of going to the hospital?
- How does a family start a suitable early intervention?

  etc.

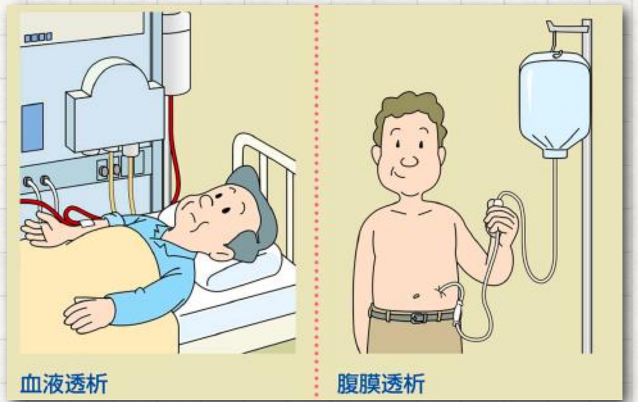
How Causal Discovery Helps Healthcare?



In 2021, the number of dialysis therapy clients was 349,700 in Japan.

2786 people per million population.

67% of them are over 65-yr-old. The average age is 69.7 yr-old.



一般社団法人
日本透析医学会
The Japanese Society for Dialysis Therapy

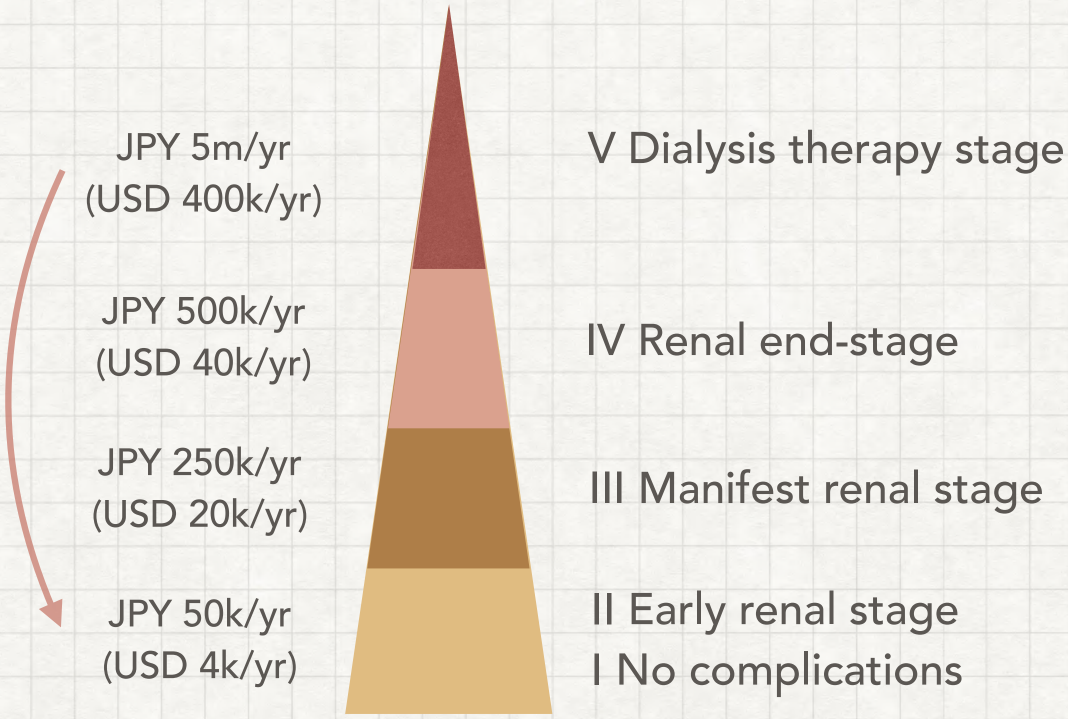
Resource: The Japanese Society for Dialysis Therapy (JSDT). 2021

Diabetic Nephropathy Severe Exacerbation Prevention Program (DNSEP Program)

For new dialysis therapy clients, 40.2% result from dialysis nephropathy. (JSDT Renal Data Registry. 2021)

JSDT: The Japanese Society for Dialysis Therapy

If one dialysis therapy client can be controlled to stay in the early stages, save the huge national medic budget, max to near JPY5m (USD400k) per year per person.



National burden portion of medical expenses per person per year. Diabetic nephropathy related.

Resource: Ministry of Health, Labor, and Welfare of Japan. 2017

Health Data Analysis by Causal Discovery for Supporting Policy Formation and Municipal Problem-Solving

（因果探索を用いた健康データ分析に基づくヘルス課題解決と政策形成支援）

(D1) トウオウ Ou DENG (dengou@toki.waseda.jp)
Graduate School of Human Sciences, Waseda University

(I) Data preparation phrase

Data sources

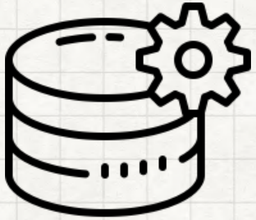
NHI(KDB)

MUNICIPAL HMS

Anonymization

Data storage

DB for analyzing



Data master
- Receipt(Treatment+Dispensing)
- Special health check-up
- NIH(KDB)
- Elderly care

Filtering

(II) Data analyzing phrase

Problems Input
- Specific diseases
- Regional health issues
- Preventive measures

Mechanism of Causal Discovery
- Time series data
- Non-time series data

Data analyzing

Visualization of Analysis Results & On-Site Evidence Verification

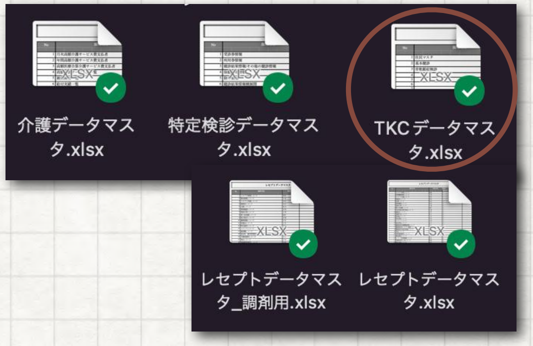
(III) Outcomes evaluation phrase

(Improvement of Outcomes)

Results output
- Causal graph of health data
- Evidence for policy formation
- Solution to municipal problems

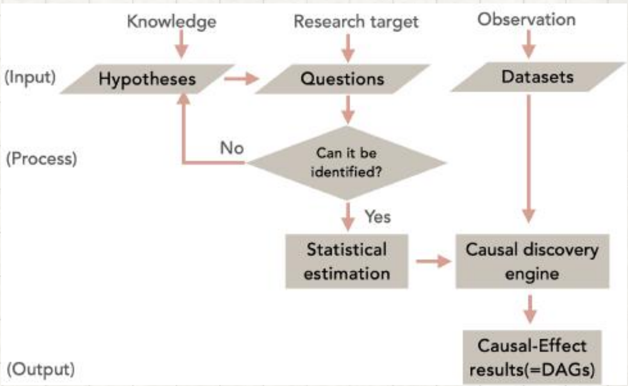
Note:
- NHI(KDB): Database system managed by National Health Insurance Organizations(NHI).
- Municipal HMS: Local Health Management System managed by the municipal government.

(I)Data preparation phrase



(II)Data analyzing phrase

Causal Discovery of Observational Multivariate



A mixed routine of classic statistics and machine learning for causal discovery

(To be continued…)

This research project is supervised by **Prof. Shoji Nishimura, Prof.Atsushi Ogihara, and Prof.Qun Jin** and supported by

- The Agreement Concerning the Regional Revitalization Project between Minano Town in Saitama Prefecture and the Faculty of Human Sciences at Waseda University. (Research coordinator: Atsushi Ogihara)
- FY2022-2024 Grants for the Promotion of the National Concept of Digital Rural Cities. (Research coordinator: Shoji Nishimura, Qun Jin)
- 2020-2025 JSPS A3 Foresight Program (Grant No. JPJSA3F20200001)