

# Klasterizavimo algoritmai didelės apimties medicinos duomenims

Doktorantas: Roma Purnaitė

Vadovas: prof. dr. Audronė Jakaitienė

Doktorantūros pradžios ir pabaigos metai: 2017-2025

2025 m. spalio 2 d.

# Išlaikyti egzaminai

- Informatikos matematiniai metodai, 9 kreditai, 2018 m. spalio 1 d.
- Duomenų analizės strategijos ir sprendimų priėmimas, 7 kreditai, 2018 m. spalio 11 d.
- Netiesiniai statistikos modeliai masinių duomenų analizėje, 7 kreditai, 2022 m. kovo 18 d.
- Daugiamačių duomenų vizualizavimo metodai, 7 kreditai, 2022 m. kovo 30 d.
- Bendrųjų gebėjimų mokymai, 3 kreditai, 2025 m. liepos 2 d.

# Publikacijos WoS (Clarivate Analytics)

1. Jurevičienė E; Onder G; Visockienė Ž; **Puronaitė R**; Petrikonytė D; Gargalskaitė U; Kasiulevičius V; Navickas R. *Does multimorbidity still remain a matter of the elderly: Lithuanian national data analysis* // Health policy : Elsevier Ireland Ltd. ISSN 0168-8510. eISSN 1872-6054. 2018, Vol. 122, no 6, p. 681-686. DOI: 10.1016/j.healthpol.2018.03.003.
2. Nedzinskienė L; Jurevičienė E; Visockienė Ž; Ulytė A; **Puronaitė R**; Kasiulevičius V; Kazėnaitė E; Burneikaitė G; Navickas R. *Structure and distribution of health care cost across age groups of patients with multimorbidity in Lithuania* // International journal of environmental research and public health. Basel : MDPI. ISSN 1661-7827. eISSN 1660-4601. 2021, vol. 18, no. 5, art. no. 2767, p. [1-13]. DOI: 10.3390/ijerph18052767.
3. Ramanauskaitė D; **Puronaitė R**; Jakaitienė A; Glaveckaitė S. *Prevalence of multimorbidity in Lithuania: insights from national health insurance fund data* // Journal of cardiovascular development and disease. Basel : MDPI. eISSN 2308-3425. 2025, vol. 12, iss. 2, art. no. 47, p. [1-18]. DOI: 10.3390/jcdd12020047.

**Indėlis:** duomenų paruošimas analizei, dauginio ligotumo paplitimo vertinimas, netiesiniai regresiniai modeliai (dauginio ligotumo paplitimas, sveikatos priežiūros resursų naudojimas), vizualizacijos, publikacijos teksto rengimo pagal kompetenciją.

# Kitos publikacijos ne WoS

1. Trinkūnas, J., Tuinylienė, E., & Purnaitė, R. (2018). Research on hospital information systems integration to national electronic health record system. In BIOMDLORE 2018: proceedings of 12th international conference, Białystok, Poland, June 28-30, 2018, Białystok, Poland (pp. 1-6). New York: IEEE. doi:10.1109/BIOMDLORE.2018.8467200.
2. Norkus, A. (sudarytojas), <...>, Kasiulevičius, V., <...>, Purnaitė, R., <...> Visockienė, Ž. ir kt. (2020). Diabetas ir kardiovaskulinė rizika: monografija. Skyrius: 2 tipo cukrinis diabetas ir poliligtumas. Kaunas: Medicininės informacijos centras.
3. Jurevičienė, E., Burneikaitė, G., Dambrauskas, L., Kasiulevičius, V., Kazėnaitė, E., Navickas, R., Purnaitė, R., Smailytė, G., Visockienė, Ž., & Danila, E. (2022). Epidemiology of chronic obstructive pulmonary disease (COPD) comorbidities in Lithuanian national database: a cluster analysis. International journal of environmental research and public health, 19(2), 1-14. doi:10.3390/ijerph19020970.

**Indėlis:** duomenų paruošimas analizei, modeliavimas, vizualizacijos, publikacijos teksto rengimo pagal kompetenciją.

# Pranešimai tarptautinėse konferencijose

- 1. Purnaitė R;** Jakaitienė A; Jurevičienė E; Kasiulevičius, V; Navickas R; Radavičius M; Visockienė Ž. *Identifying patterns of multimorbidity in Lithuanian National Health Insurance Fund data: a comparison of cross-sectional and temporal phenotyping approaches* // NBBC19 : 7th Nordic-Baltic biometric conference, 3-5 June 2019, Vilnius, Lithuania.
- 2. Purnaitė R;** Jakaitienė A; Švaikevičienė K; Burneikaitė G; Trinkūnas J; Kasiulevičius, V; Kazėnaitė E. *Challenges of using big health data to identify patterns of anxiety and depression in multimorbid population* // ISCB 2021: 42nd annual conference of the International Society For Clinical Biostatistics, 18-22 July 2021, Lyon, France.
- 3. Purnaitė R;** Ramanauskaitė D; Burneikaitė G; Švaikevičienė K; Šavareikaitė A; Vaitkutė S; Jakaitienė A; Dambrauskas L; Jurevičienė E; Trinkūnas J; Kasiulevičius V; Kazėnaitė E. *Challenges of modeling depression and anxiety risk using data from large healthcare databases: systematic review and situation analysis* // 31st International Biometric Conference (IBC2022), 10 - 15 July 2022, Riga, Latvia
- 4. Purnaitė R;** Ramanauskaitė D; Švaikevičienė K; Tarutyte G; Trinkūnas J; Burneikaitė G; Kazėnaitė E; Glaveckaitė S; Jakaitienė A. *Using nonlinear models to identify breakpoints in disease prevalence among patients with multimorbidity* // ISCB44: 44th annual conference of the International Society for Clinical Biostatistics; 27-31 August 2023, Milan, Italy.
- 5. Purnaitė R;** Tarutyte G; Ramanauskaitė D; Burneikaitė G; Glaveckaitė S; Jakaitienė A. *Impact of methodological choices on multimorbidity prevalence estimates in Lithuania* // 1st international Vilnius conference on Statistics and its applications, August 27-29, 2025, Vilnius, Lithuania.

# Tyrimo objektas, tikslas ir uždaviniai

## **Tyrimo objektas**

Dauginio ligotumo paplitimo ir pasekmių sveikatos politikai modeliavimas.

## **Tyrimo tikslas**

Optimalus dauginio ligotumo konceptą atitinkančių kriterijų parinkimas ir dauginio ligotumo netiesinis modeliavimas panaudojant nacionalinius administracinius sveikatos duomenis.

## **Uždaviniai**

1. Pasiūlyti dauginio ligotumo Lietuvos populiacijos duomenų paruošimo modeliavimui metodiką.
2. Pasiūlyti modelius dauginio ligotumo rodiklių analizei svarbius Lietuvos sveikatos politikai.
3. Atlikti eksperimentinę statistinę duomenų analizę, naudojant netiesinius statistinius metodus nacionaliniams administraciniams sveikatos duomenims.
4. Daugiakriterinis sprendimų priėmimas optimaliai dauginio ligotumo koncepto empirinei realizacijai įvertinti.

# Tyrimo metodai ir duomenys

## Metodai

- Netiesiniai statistikos metodai.
- Daugiakriterinis sprendimų priėmimas.
- Klasterizavimo algoritmai.

## Naudoti duomenys

- Lietuvos populiacijos administraciniai sveikatos duomenų rinkiniai: Valstybinės ligonių kasos valstybinio sveikatos draudimo lėšomis apmokamų sveikatos priežiūros paslaugų apskaitos duomenų bazės (SVEIDRA) duomenų rinkiniai:
  - 2012-2014 metų (pacientų skaičius: apie 450 tūkst., vizitų skaičius: apie 17 mln., diagnozių įrašų skaičius: apie 30 mln.)
  - 2014-2019 metų (pacientų skaičius: apie 1,8 mln., vizitų skaičius: apie 170 mln., diagnozių įrašų skaičius: apie 60 mln.)
- Valstybės duomenų agentūros informacija apie gyventojų skaičių kiekvienų metų pradžioje.

# Dauginis ligotumas

- Dauginis ligotumas (angl. *multimorbidity*) – tai medicinos ir visuomenės sveikatos konceptas, apibūdinanti būklę, kai vienas asmuo serga keliomis lėtinėmis ligomis vienu metu.
- Būklė siejama su didesniu krūviu sveikatos sistemai, prastesne gyvenimo kokybe, komplikuotais ligų ryšiais.

## **Vertinimo strategijos:**

- Skirtingi būklės apibrėžimai.
- Administraciniai duomenų rinkiniai.

## **Iššūkiai:**

- Nėra vieningo apibrėžimo.
- Nėra sutarimo kaip vertinti naudojant administracinius sveikatos duomenis.
- Skirtingų šalių duomenų bazių struktūrų įvairovė.



# Siūloma dauginio ligotumo vertinimo metodika

## 1. Administracinių duomenų parinkimas.

- Kiek įmanoma pilnesni (laiko intervalas, paslaugų apimtis).

## 2. Paruošimas.

- Klinikinė dalis – ligų atranka, diagnozių sąrašai, paslaugų filtravimas.
- Informatikos dalis – ETL procesas, filtravimas, ligų identifikavimas rinkinyje, alternatyvių pasirinkimų sudarymas ir atranka.

## 3. Dauginio ligotumo koncepto empirinės realizacijos vertinimas.

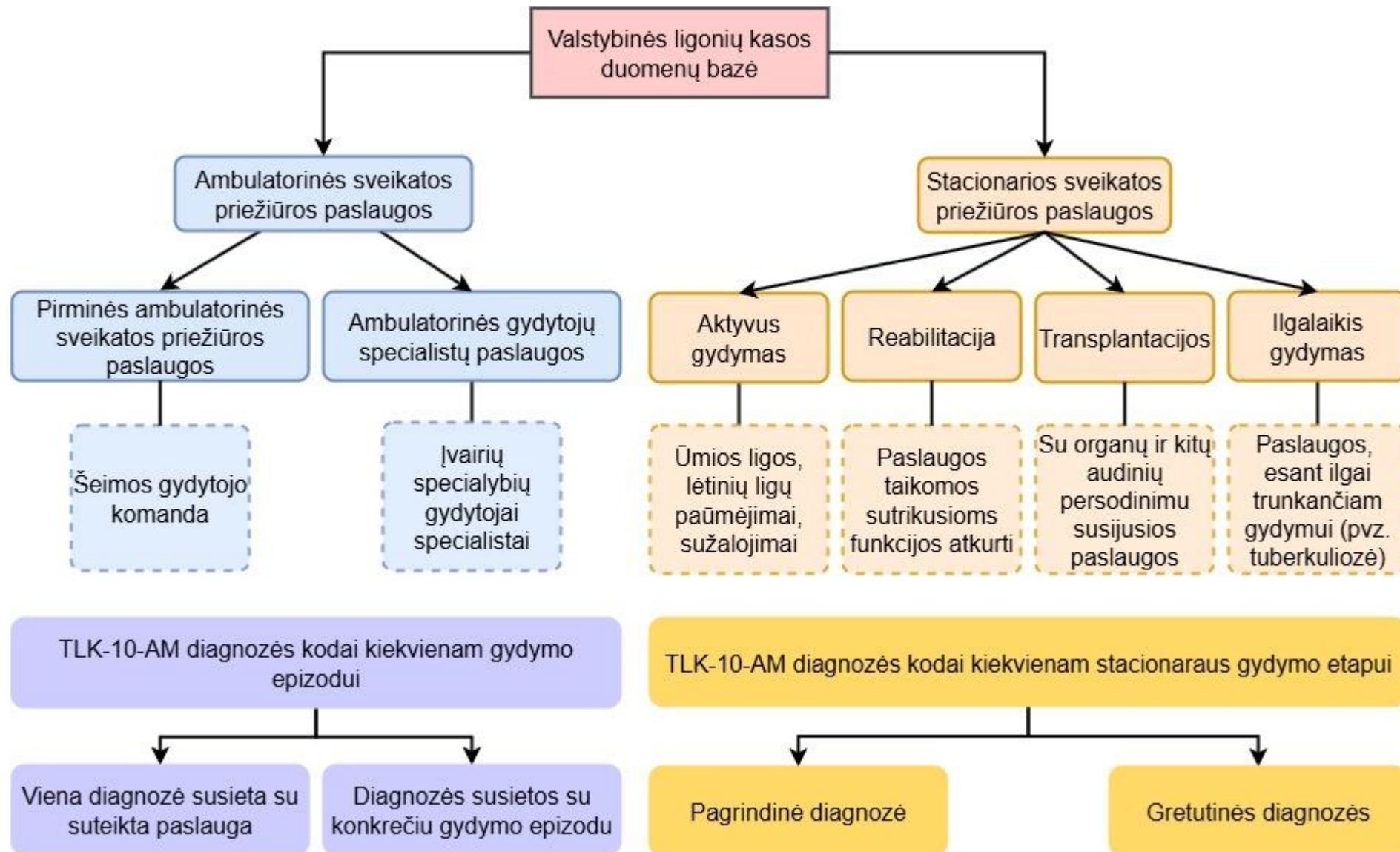
- Alternatyvūs pasirinkimai: duomenų šaltinis, laiko intervalas, diagnozės atsikartojimai, apibrėžimas.
- Apibrėžiami kriterijai atitinkantys dauginio ligotumo konceptas (literatūros analizė, ankstesni tyrimai).
- Alternatyvų vertinimas daugiakriteriais sprendimų priėmimo metodais.

# Siūloma dauginio ligotumo vertinimo metodika

## 4. Galutinio premdimo atranka.

- Alternatyvių pasirinkimų rangavimas ir klasterizavimas.
- Geriausiai įvertintų alternatyvų variabilumo vertinimas.
- Pasirinkimas: tiksliausiai dauginio ligotumo konceptą atitinkanti alternatyva ar kita alternatyva iš geriausiai įvertintų, atsižvelgiant į variabilumą.

# Duomenų bazės struktūra



# Dauginio ligotumo koncepto empirinės realizacijos vertinimas

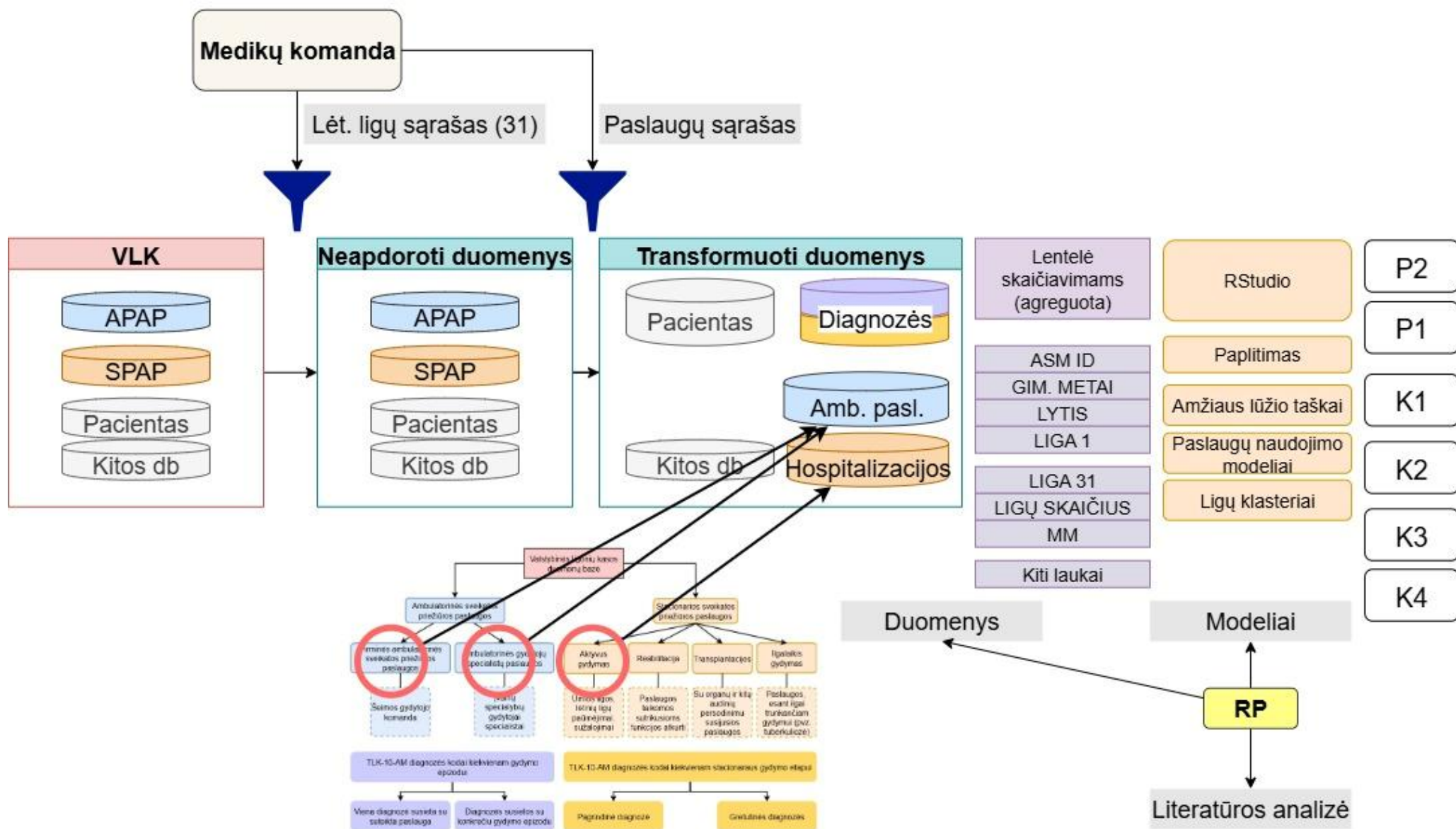
- 31 lėtinė liga (diagnozių rinkiniai).
- 630 alternatyvių dauginio ligotumo įvertinimo pasirinkimų (šaltinis, laiko intervalas, diagnozės atsikartojimas ligai patvirtinti, dauginio ligotumo apibrėžimas).
- 20 papildomų alternatyvų individualiai ligai parenkant diagnozės atsikartojimų skaičių ligai patvirtinti.
- Kriterijai: (1) aukšta santykinė didelio sveikatos paslaugų naudojimo intensyvumo rizika, (2) aukšta santykinė mirties rizika, (3) aukštas vidutinis Jaccard'o koeficientas.

# Galimi metodologiniai pasirinkimai duomenų paruošimo etape

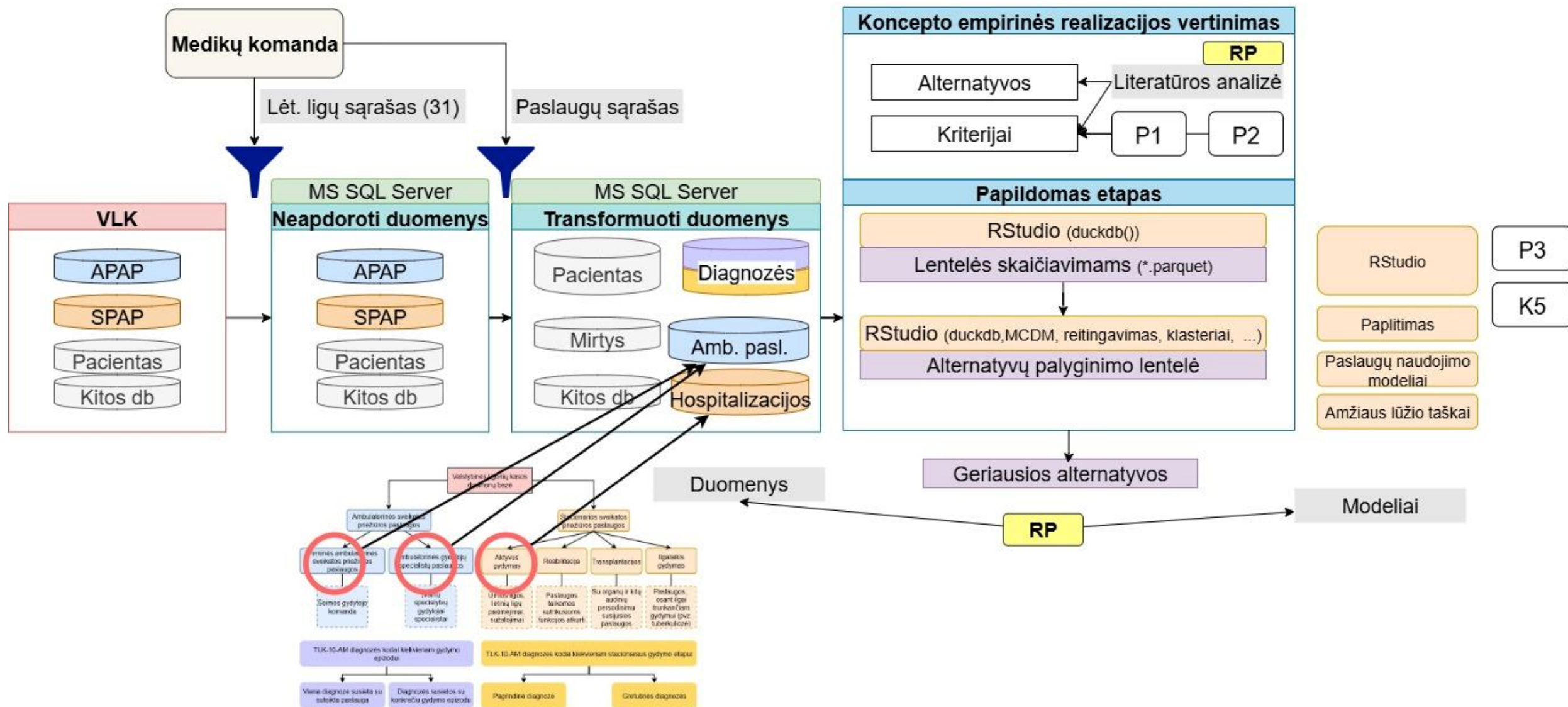
| Duomenų šaltinis                                                                                                                                              | Laiko intervalas                                                                                                                                                                    | Ligos patvirtinimui pasirinktas atsikartojimo DB skaičius                                                                                                      | Dauginio ligotumo apibrėžimas                                                                                                                                                                                                                                       |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>S1: Šeimos medicina (P)</p> <p>S2: Specialistai (O)</p> <p>S3: Stacionaras (I)</p> <p>S4: P + O</p> <p>S5: P + I</p> <p>S6: O + I</p> <p>S7: P + O + I</p> | <p>T1: Tik analizuojami metai*</p> <p>T2: +1 metai atgal**</p> <p>T3: +2 metai atgal</p> <p>T4: +3 metai atgal</p> <p>T5: +4 metai atgal</p> <p>T5: +5 metai atgal</p>              | <p>N1: Bent 1 kartą per analizuojamą laikotarpį*</p> <p>N2: Bent 2 kartus per analizuojamą laikotarpį</p> <p>N3: Bent 3 kartus per analizuojamą laikotarpį</p> | <p>M1: Bent dvi lėtinės ligos*</p> <p>M2: Bent trys lėtinės ligos</p> <p>M3: Bent dvi lėtinės ligos iš bent dviejų specialybių**</p> <p>M4: Bent trys lėtinės ligos iš bent dviejų specialybių</p> <p>M5: Bent viena fizinė ir bent viena psichiatrinė diagnozė</p> |
|                                                                                                                                                               | <p>*pvz., tikslas įvertinti dauginio ligotumo būklę 2019 metais</p> <p>** tada įvertinsime paplitimą 2019 metais, remdamiesi tuo ar asmuo gydėsi dėl lėt. ligų 2018-2019 metais</p> | <p>* Pasirinktam šaltiniui ir vertinimo laikotarpiui</p>                                                                                                       | <p>*pvz., hipertenzija ir išeminė širdies liga</p> <p>** pvz., hipertenzija ir cukrinis diabetas (kardiologija, endokrinologija)</p>                                                                                                                                |
| MacRae et al. (2023A), Ho et al. (2021)                                                                                                                       | Simard et al. (2024)                                                                                                                                                                | Simard et al. (2024), Beaney et al. (2024)                                                                                                                     | Simard et al. (2024), MacRae et al. (2023B), Ho et al. (2021), Chua et al. (2021)                                                                                                                                                                                   |

I etapas

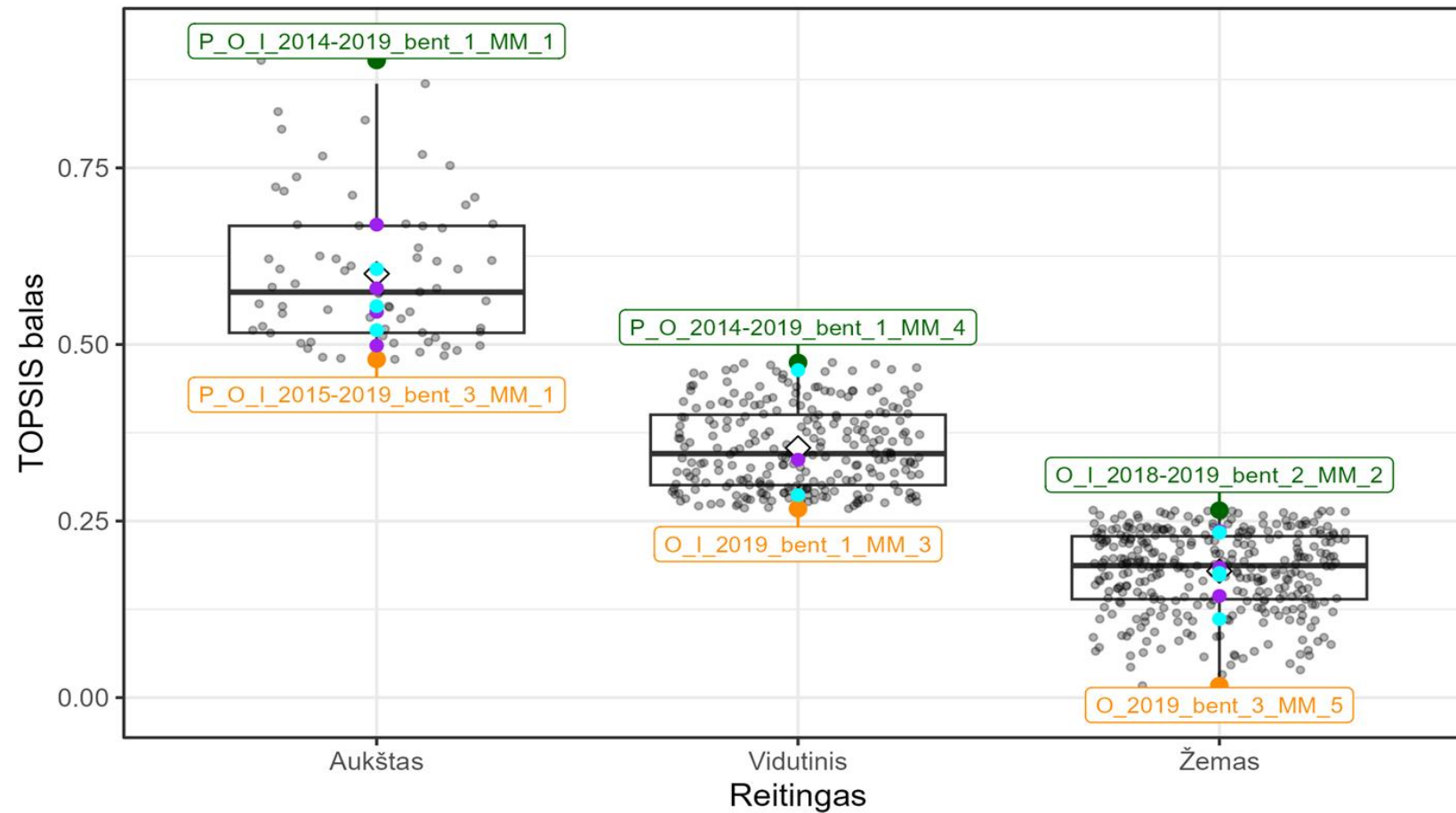
Duomenys  
2012-2014







# Rangavimo rezultatai



Individualus slenksčio parinkimas ligos patvirtinimui    Metodas: mediana    Metodas: moda



# Tyrimo eiga

## Netiesinis modeliavimas publikacijose

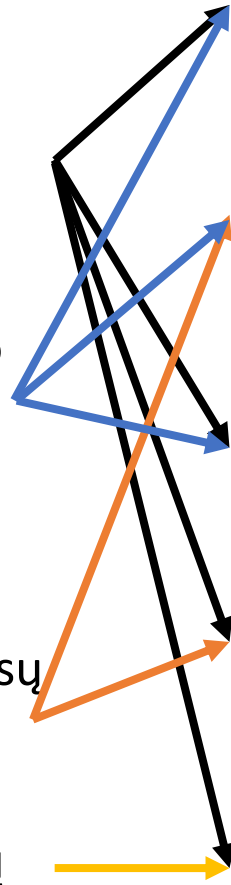
Sveikatos politikai reikšmingi dauginio ligotumo aspektai:

1) paplitimo vertinimas,

1) amžiaus lūžio taškų, kuriuose pakinta dauginio ligotumo paplitimas amžiaus atžvilgiu, identifikavimas,

1) sveikatos priežiūros resursų naudojimo modeliai,

1) metodologinių sprendimų priėmimas duomenų paruošimo procese.



1. Jurevičienė, Onder, Visockienė, **Puronaitė** et al. Does multimorbidity still remain a matter of the elderly: Lithuanian national data analysis // Health policy (2018) DOI: 10.1016/j.healthpol.2018.03.003.

2. Nedzinskienė, Jurevičienė, Visockienė, Ulytė, **Puronaitė** et al. Structure and distribution of health care cost across age groups of patients with multimorbidity in Lithuania // International journal of environmental research and public health. (2021) DOI: 10.3390/ijerph18052767

3. **Puronaitė**; Ramanauskaitė, Švaikevičienė, et al. Using nonlinear models to identify breakpoints in disease prevalence among patients with multimorbidity // ISCB 2023.

4. Ramanauskaitė, **Puronaitė**, Jakaitienė, Glaveckaitė. Prevalence of multimorbidity in Lithuania: insights from national health insurance fund data (2025) DOI: 10.3390/jcdd12020047.

5. **Puronaitė**, Tarutytė, Ramanauskaitė et al. Impact of methodological choices on multimorbidity prevalence estimates in Lithuania // ICSA 2025

# Tyrimo eiga

## Netiesinis modeliavimas metodai

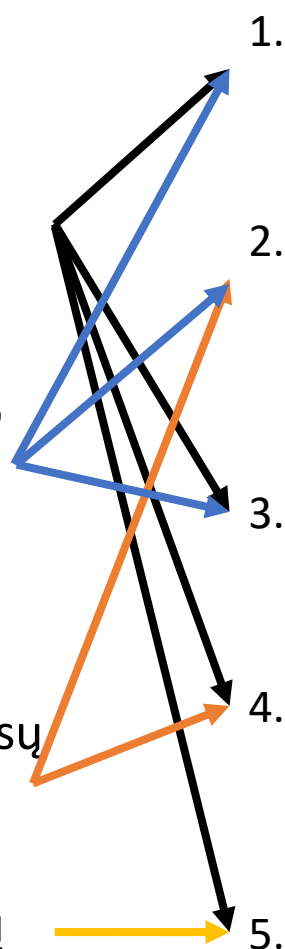
Sveikatos politikai reikšmingi dauginio ligotumo aspektai:

1) paplitimo vertinimas,

1) amžiaus lūžio taškų, kuriuose pakinta dauginio ligotumo paplitimas amžiaus atžvilgiu, identifikavimas,

1) sveikatos priežiūros resursų naudojimo modeliai,

1) metodologinių sprendimų priėmimas duomenų paruošimo procese.



1. Segmentinė regresija

2. Segmentinė regresija, apibendrintieji tiesiniai modeliai

3. Segmentinė regresija, klasifikavimo ir regresijos medžiai, adityvieji regresijos modeliai

4. Perteklinių nulių modeliai, barjeriniai modeliai

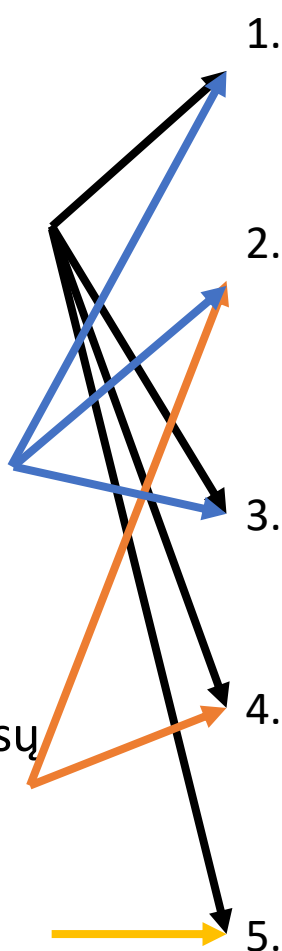
5. Daugiakriteriniai sprendimų priėmimo metodai, klasterizavimas

# Tyrimo eiga

## Netiesinis modeliavimas rezultatai

Sveikatos politikai reikšmingi dauginio ligotumo aspektai:

- 1) paplitimo vertinimas,
- 1) amžiaus lūžio taškų, kuriuose pakinta dauginio ligotumo paplitimas amžiaus atžvilgiu, identifikavimas,
- 1) sveikatos priežiūros resursų naudojimo modeliai,
- 1) metodologinių sprendimų priėmimas duomenų paruošimo procese.



1. Pritaikius segmentinę regresijos modelį dauginio ligotumo amžiaus taškuose modeliavimui gauti amžiaus lūžio taškai, kuriuose pasikeičia dauginio ligotumo paplitimas populiacijoje.

2. Modeliuojant sveikatos sistemos išlaidas nustatyta, kad geriausias buvo gama regresijos modelis su logaritmine jungties funkcija. Naudojant segmentinę regresiją identifikuoti amžiaus lūžio taškau, kuriuose pakinta išlaidos tenkančios žmogui per metus.

3. Segmentinės regresijos modeliai buvo tiksliausi modeliuojant konkrečių ligų turėjimą dauginėmis ligomis sergantiems pacientams, tačiau testavimo imtyje aukštas tikslumas pasiekiamas ne visoms ligoms.

4. Platesnės apimties duomenų rinkinys pritaikytas paplitimui įvertinti ir sveikatos priežiūros resursų naudojimui modeliuoti. Geriausias tikslumas pasiektas naudojant barjerinę (angl. *hurdle*) neigiamąją binominę regresiją.

5. Nustatyta, kad alternatyvūs apibrėžimai keičia paplitimo įverčius. Itin platūs intervalai 80+ amžiaus grupėse. Lietuvos atveju geriausia alternatyva – maksimali apimtis ir laikotarpis, bent viena diagnozė ligai, bent dvi ligos dauginiam ligotumui apibrėžti. Kitos aukšto reitingo alternatyvos leidžia mažinti laiko intervalą ar atsikartojimų skaičių, bet ne duomenų šaltinį.

# Išvados

1. Tyrimo metu parinkti ir pritaikyti modeliai sveikatos politikai reikšmingų dauginio ligotumo rodiklių analizei: krūvis ir kaina sveikatos sistemai, paplitimas ir amžiaus lūžio taškai paplitimo pokyčiams.
2. Buvo atlikta statistinė analizė, taikant netiesinius statistinius metodus Lietuvos administraciniams sveikatos duomenims.
3. Pasiūlytas sprendimų priėmimo modelis dauginio ligotumo koncepto empirinės realizacijos vertinimui pritaikytas Lietuvos sveikatos sistemos struktūrai.
4. Taikytas reitingavimo principas leidžia įvertinti, kaip kiekvienas metodologinis pasirinkimas prisideda prie galutinio rezultato.
5. Kitose populiacijose tikslinga analogiškai ištirti, koks duomenų rinkinys ir dauginio ligotumo vertinimo metodas geriausiai atitinka dauginio ligotumo konceptą.
6. Reikėtų sutarti dėl kriterijų rinkinio, kuris leistų užtikrinti, kad kiekvieno konkretaus tyrimo realizacija maksimaliai priartėtų prie apibrėžto optimalaus rezultato.

# Literatūra

M. Simard, E. Rahme, M. Dubé, V. Boiteau, D. Talbot, and C. Sirois, 'Multimorbidity prevalence and health outcome prediction: assessing the impact of lookback periods, disease count, and definition criteria in health administrative data at the population-based level', *BMC Med Res Methodol*, vol. 24, no. 1, Art. no. 1, May 2024, doi: [10.1186/s12874-024-02243-0](https://doi.org/10.1186/s12874-024-02243-0).

T. Beaney, J. Clarke, T. Woodcock, A. Majeed, M. Barahona, and P. Aylin, 'Effect of timeframes to define long term conditions and sociodemographic factors on prevalence of multimorbidity using disease code frequency in primary care electronic health records: retrospective study', *bmjmed*, vol. 3, no. 1, Art. no. 1, Feb. 2024, doi: [10.1136/bmjmed-2022-000474](https://doi.org/10.1136/bmjmed-2022-000474).

C. MacRae *et al.*, 'Impact of data source choice on multimorbidity measurement: a comparison study of 2.3 million individuals in the Welsh National Health Service', *BMC Med*, vol. 21, no. 1, Art. no. 1, Aug. 2023, doi: [10.1186/s12916-023-02970-z](https://doi.org/10.1186/s12916-023-02970-z). (A)

C. MacRae *et al.*, 'Age, sex, and socioeconomic differences in multimorbidity measured in four ways: UK primary care cross-sectional analysis', *Br J Gen Pract*, vol. 73, no. 729, Art. no. 729, Apr. 2023, doi: [10.3399/BJGP.2022.0405](https://doi.org/10.3399/BJGP.2022.0405). (B)

I. S.-S. Ho *et al.*, 'Examining variation in the measurement of multimorbidity in research: a systematic review of 566 studies', *The Lancet Public Health*, vol. 6, no. 8, Art. no. 8, Aug. 2021, doi: [10.1016/S2468-2667\(21\)00107-9](https://doi.org/10.1016/S2468-2667(21)00107-9).

Y. P. Chua, Y. Xie, P. S. S. Lee, and E. S. Lee, 'Definitions and Prevalence of Multimorbidity in Large Database Studies: A Scoping Review', *IJERPH*, vol. 18, no. 4, Art. no. 4, Feb. 2021, doi: [10.3390/ijerph18041673](https://doi.org/10.3390/ijerph18041673).