Klasterizavimo algoritmai didelės apimties medicinos duomenims

Ataskaita už II-ąjį doktorantūros kursą

Doktorantas: Roma Puronaitė
Vadovas: prof. dr. Audronė Jakaitienė

Doktorantūros pradžios ir pabaigos metai: 2017 - 2023

2021 m. rugsėjis

Tyrimo objektas

- Klasterizavimo algoritmai
- Didelės apimties medicinos duomenys

Tyrimo tikslas

- Pasiūlyti metodą didelės apimties medicinos duomenims klasterizuoti, atsižvelgiant į duomenų dinamikos laike savybes.

Uždaviniai

1. Ištirti didelės apimties medicinos duomenų klasterizavimui dažniausiai taikomus klasterizavimo metodus.
2. Pasiūlyti klasterizavimo algoritmą ar esamo metodo patobulinimą, kuris atsižvelgtų į dinamiką laike.
3. Pritaikyti atrinktus algoritmus ir pasiūlytą sprendimą realiems medicinos duomenų rinkiniams.
4. Pasiūlyti algoritmo integravimo į sveikatos priežiūros įstaigos informacinę sistemą modelį.

Doktorantūros planas

Studijų metai	Egzaminai		Dalyvavimas konferencijose		Publikacijos		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė [4]
I (2017/2018)	2	2		1		1	1 publikuota
II (2018/2019)	2	0	1	5		2+1	2 publikuota, 1 Įteikta (gautos pirmos recenzijos)
III (2019/2020) **III (2021/2022)**	2 (skola iš II metų**)		1		1		
IV (2020/2021) **IV (2022/2023)**					*1*		
Iš viso:	4	2	2	6	2	3+1	

*Iki 2021 m. rugsėjo 30 d. akademinės atostogos
**Prašau leisti koreguoti doktorantūros planą ir leisti du II metais numatytus egzaminus išlaikyti per III metus.

1 pav.: Visų studijų planas

Ataskaita už II mokslo metus

1. Egzaminai:
   - Daugiamačių duomenų vizualizavimo metodai (prašoma perkelti į III metų I pusm.).
   - Netiesiniai statistikos modeliai masinių duomenų analizėje (prašoma perkelti į III metų II pusm.).

Ataskaita už II mokslo metus

1 Publikacijos:

Norkus, Antanas (sudaryt.); Kasiulevičius, Vytautas; **Puronaitė, Roma**; Visockienė, Žydrūnė <...> **Diabetas ir kardiovaskulinė rizika : monografija – skyrius: 2 tipo cukrinis diabetas ir poliligotumas** / sudarytojas Antanas Norkus. Kaunas : Medicininės informacijos centras, 2020. 386 p. ISBN 9786098070279.	Parengtas skyrius monografijoje, kuriame pateikti klasterizavimo metodų taikymo rezultatai poliligotiems pacientams.
Nedzinskienė, Laura; Jurevičienė, Elena; Visockienė, Žydrūnė; Ulytė, Agnė; **Puronaitė, Roma**; Kasiulevičius, Vytautas; Kazėnaitė, Edita; Burneikaitė, Greta; Navickas, Rokas. **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.	Parengta publikacija nagrinėjant didelės apimties duomenų sveikatos duomenų rinkinį, naudojant segmentinės regresijos metodą identifikuoti lūžio taškai (pacientų amžius), kuriuose pasikeičia ligų skaičiaus ir išlaidų dydžio trajektorijos.
Antanas Bliudzius, **Roma Puronaite**, Justas Trinkunas, Audrone Jakaitiene, Vytautas Kasiulevicius **Research on physical activity variability and changes of metabolic profile in patients with prediabetes using Fitbit activity trackers data** (įteiktas po recenzijos) bus spausdinamas (09.24 gautas patvirtinimas)	Įteiktas žurnalui: Technology and Health Care (IF: 1.285) Parengta publikacija pritaikant Poincare plot metodą ir Poincare indeksus, fizinio aktyvumo variabilumo vertinimui susiejant su laboratoriniais rodikliais ir kūno sudėties ištyrimo rodikliais. Fizinio aktyvumo variabilumui įvertinti šis metodas pritaikomas pirmą kartą.

2 pav.: Publikacijos

Ataskaita už II mokslo metus

Dalyvavimas konferencijose:

Puronaitė, Roma; Jakaitienė, Audronė; Jurevičienė, Elena; Kasiulevičius, Vytautas; Navickas, Rokas; Radavičius, Marijus; Visockienė, Žydrūnė. **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 : final programme and abstract book. Vilnius : [s.n.]. 2019, p. 39.	Pristatė R. Puronaitė. Žodiniame pranešime pristatyti rezultatai gauti cukriniu diabetu sergančių pacientų grupei naudojant administracinio pobūdžio poliligotų pacientų sveikatos duomenų bazę pritaikius dažniausiai šioje srityje taikomus klasterizavimo metodus hierarchinį klasterizavimą ir faktorinę analizę ir išmėginus į kaitą laike atsižvelgiantį metodą PARAFAC2.
Roma Puronaitė, Audronė Jakaitienė, Kristina Švaikevičienė, Greta Burneikaitė, Justas Trinkūnas, Vytautas Kasiulevičius and Edita Kazėnaitė **Challenges of using big health data to identify patterns of anxiety and depression in multimorbid population** 42nd Conference of the International Society for Clinical Biostatistics (ISCB) 2021	Pristatė R. Puronaitė. Pranešime pristatyti iššūkiai su kuriais susiduriama analizuojant didelės apimties administracinio pobūdžio sveikatos duomenų bazes. Pristatyti klasterizavimo metodų pritaikymo depresijos ir nerimo diagnozes turintiems pacientams rezultatai.
Roma Puronaitė, Kristina Švaikevičienė, Greta Burneikaitė, Dovilė Ramanauskaitė, Justas Trinkūnas, Vytautas Kasiulevičius, Audronė Jakaitienė, Edita Kazėnaitė **Identification of Anxiety and Depression Patterns in Multimorbid Patients: A Case of Secondary Use of Administrative Health Data** Life Sciences Baltics 2021, rugsėjo 22-24 d.	Pristatė R. Puronaitė. Pranešime pristatomas didelės apimties administracinio pobūdžio sveikatos duomenų bazės panaudojimo antriniu tikslu pavyzdys. Pristatomi klasterizavimo algoritmų taikymo rezultatai.

3 pav.: Pristatyti pranešimai

Ataskaita už II mokslo metus

Dalyvavimas konferencijose (bendraautorystė):

Jurevičienė, Elena; **Puronaitė, Roma;** Danila, Edvardas. **The impact of multimorbidity on hospitalizations in patients with chronic obstructive pulmonary disease (COPD): a population-based study** // European respiratory journal: vol. 54, suppl. 63: ERS International Congress 2019 abstracts. Sheffield : European respiratory society. ISSN 0903-1936. eISSN 1399-3003. 2019, vol. 54, suppl. 63, abstract no. PA4296, p. [1]. DOI: 10.1183/13993003.congress-2019.PA4296.

Jurevičienė, Elena; Puronaitė**, Roma; Sudmantaitė,** Vaida; Danila, Edvardas. **The association of diuretics treatment with hospitalization rates of patients with chronic obstructive pulmonary disease and pulmonary hypertension: a population-based study** // European respiratory journal. Sheffield : European Respiratory Society. ISSN 0903-1936. eISSN 1399-3003. 2020, vol. 56, suppl. 64, p. 301. DOI: 10.1183/13993003.congress-2020.301.

Pranešimus pristatė bendraautorė. Atlikta skerspjūvio tipo duomenų analizė, remiantis administracinio pobūdžio duomenų baze įvertintas gydymo nuoseklumo, kaip vartojamų receptinių vaistų pertrūkių vertinimas.

4 pav.: Bendraautorių pristatyti pranešimai

Ataskaita už II mokslo metus

Dalyvavimas konferencijose (bendraautorystė):

Visockienė, Žydrūnė; **Puronaitė, Roma**; Navickas, Rokas; Jurevičienė, Elena; Kasiulevičius, Vytautas. **Risk of hospitalization for cardiovascular events associated with diabetes therapies in type 2 diabetes patients in Lithuanian cohort** // Endocrine abstracts: vol. 70: 22nd European Congress of Endocrinology, 5-9 September 2020. Bristol : BioScientifica. ISSN 1470-3947. eISSN 1479-6848. 2020. vol. 70, abstract no. EP189, p. [1]. DOI: 10.1530/endoabs.70.EP189.	Pranešimą pristatė bendraautorė. Atlikta statistinė duomenų analizė, sukonstruotas algoritmas gydymui ir jo tipui įvertinti administracinio pobūdžio duomenų bazėje, neturint klinikinių duomenų ir tikslios informacijos apie gydymo pradžią.

5 pav.: Bendraautorių pristatyti pranešimai

Ataskaita už II mokslo metus

Atlikta sisteminė literatūros apžvalga Iš 162 šaltinių perskaičius pilną tekstą atmesta 61, likę panaudoti informacijai sisteminti. Identifikuoti šie tematikoje naudojami klasterizavimo metodai:

- Hierarchinis klasterizavimas (angl. HCA) (30 šaltinių)
- Latentinių klasių analizė (angl. LCA) (28 šaltiniai)
- k-vidurkių metodas (angl. k-means) (13 šaltinių)
- Tiriamoji faktorinė analizė (angl. EFA) (12 šaltinių)
- Tinklinės analizės metodas (angl. Network analysis, NetAnal) (6 šaltiniai)
- Dauginė atitikties analizė (angl. Multiple multiple correspondence analysis, MCA) (6 šaltiniai)
- Klasifikavimo ir regresijos medžiai (angl. CART) (2 šaltiniai)
- Fuzzy c-means (2 šaltiniai)

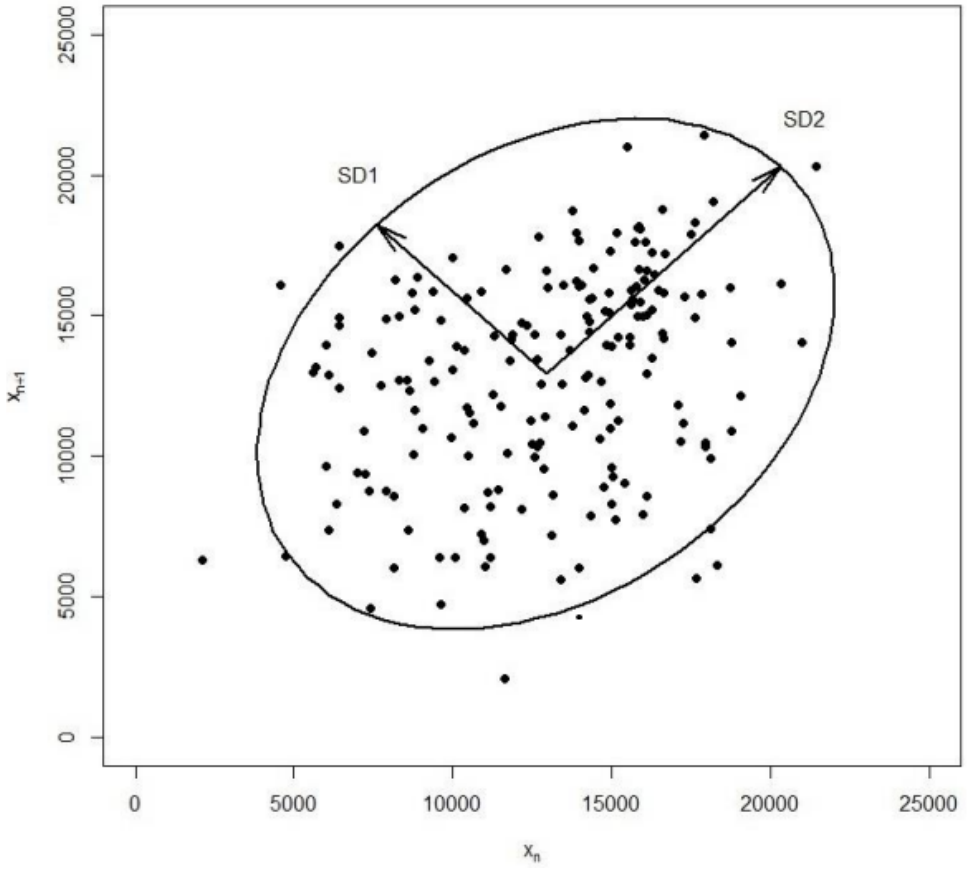
Ataskaita už II mokslo metus

Fenotipavimo laike metodai (angl. Temporal phenotyping) (4 šaltiniai):
- Paslėptasis Markovo modelis (angl. hidden Markov model, HMM),
- Tenzorių faktorizacija (angl. Tensor factorization),
- Neneigiama matricos faktorizacija (angl. Non-negative Matrix factorization, NNMF) ,
- Dinaminis laiko skalės kraipymas, Dynamic Time Warping, DTW

Ataskaita už II mokslo metus

1. Identifikuota problematika:
   - Dažniausiai atliekama tik skerspjūvio tipo analizė.
   - Neįtraukiami kintantys duomenys.
2. Siūlymai:
   - Duomenims apibūdinti naudoti papildomą informaciją, pvz. įvertinti kintamumą (angl. variability).
   - Pritaikyti Poincare plot metodą didelės apimties medicinos duomenims.

Ataskaita už II mokslo metus



$$SD1 = \frac{\sqrt{2}}{2} * SD(x_n - x_{n+1})$$

$$SD2 = \sqrt{2SD(x_n)^2 - \frac{1}{2}SD(x_n - x_{n+1})^2}$$

$$SD12 = \frac{SD1}{SD2}$$

$$AFE = \pi * SD1 * SD2$$

6 pav.: Poincare plot metodas

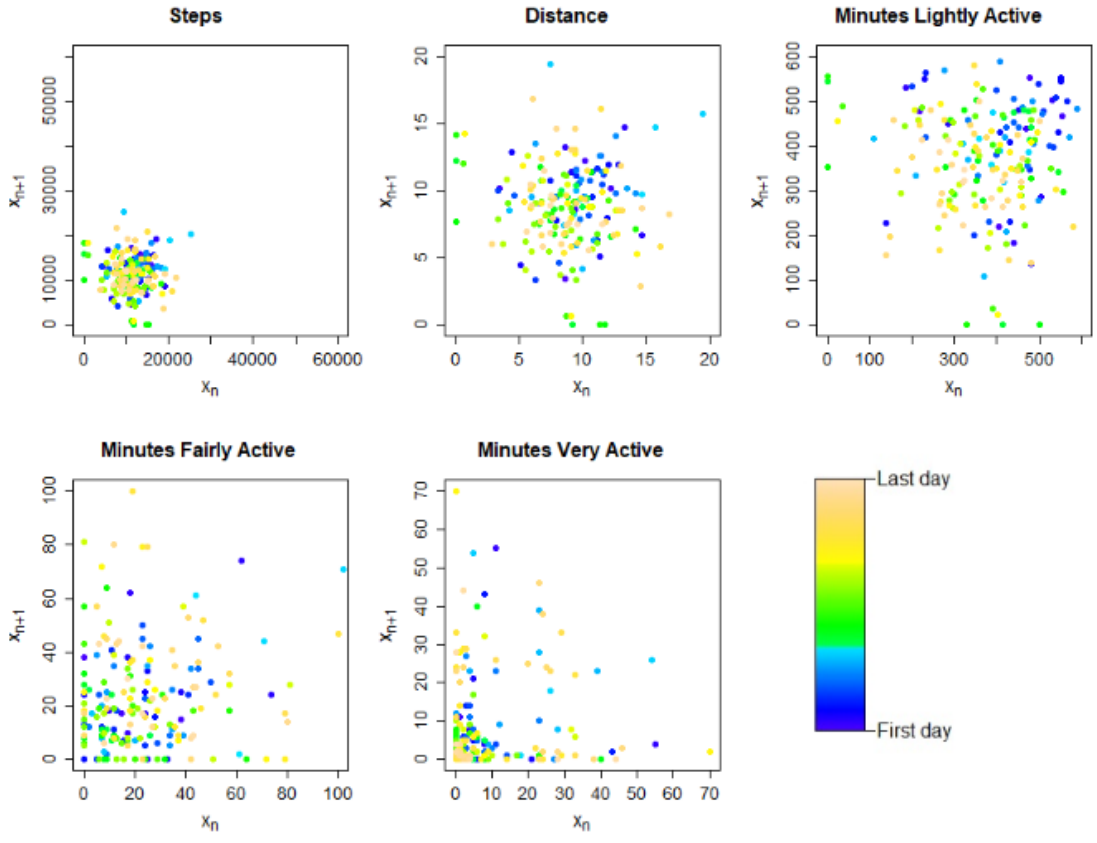
Ataskaita už II mokslo metus

Duomenų rinkiniai:

- Poliligotumo duomenų rinkinys
  - Duomenų rinkinį sudaro 1 254 167 poliligotų (turinčių dvi ar daugiau lėtinių ligų) pacientų duomenys, padengiantys laikotarpį nuo 2014 iki 2019 metų.
  - Administracinė medicinos duomenų bazė (hospitalizacijos, ambulatoriniai vizitai, diagnozės, vaistai, intervencijos)
- COVID-19 EHR
  - COVID-19 EHR duomenų bazę sudaro pacientų gydytų Všį Vilniaus universiteto ligoninės Santaros klinikose (VULSK) ir turėjusių COVID-19 ligos diagnozę (pagal TLK-10 klasifikaciją: U07.1) duomenys.
  - Duomenų rinkinį sudaro 8686 pacientų duomenys, iš jų 3009 gydytų stacionare, padengiantys laikotarpį nuo 2020-01-01 iki 2021-03-31.
  - Administraciniai duomenys + klinikiniai duomenys
- Fizinio aktyvumo duomenų rinkiniai: 30 prediabeto tyrimo pacientų duomenys, 124 personalizuoto fizinio aktyvumo tyrimo pacientų duomenys.

Ataskaita už II mokslo metus

Eksperimentai:



7 pav.: Poincare plot metodas: fizinio aktyvimo prediabeto duomenys, vieno paciento fizinio aktyvumo duomenų pavyzdys

Ataskaita už II mokslo metus

1 Eksperimentai:

Numeris 1

- 60-69 metų amžiaus grupė
- Vyras
- Lovadieniai: 35
- RITS
- Gydymo rezultatas: Mirė
- Diagnozės: Širdies nepakankamumas, kvėpavimo nepakankamumas, Sepsis, Septinis šokas, Pneumonija, COVID-19
- Laboratoriniai tyr.:

CRB (mg/l)	eGFR (CKD-EPI) (mL/min/ 1.73 m^2)	IL-6 Interleukinas-6 (ng/l)	K (mmol/l)	Na (mmol/l)	Šlapalas (mmol/l)
125	90	33,8	4,5	133	4,4

Numeris 2

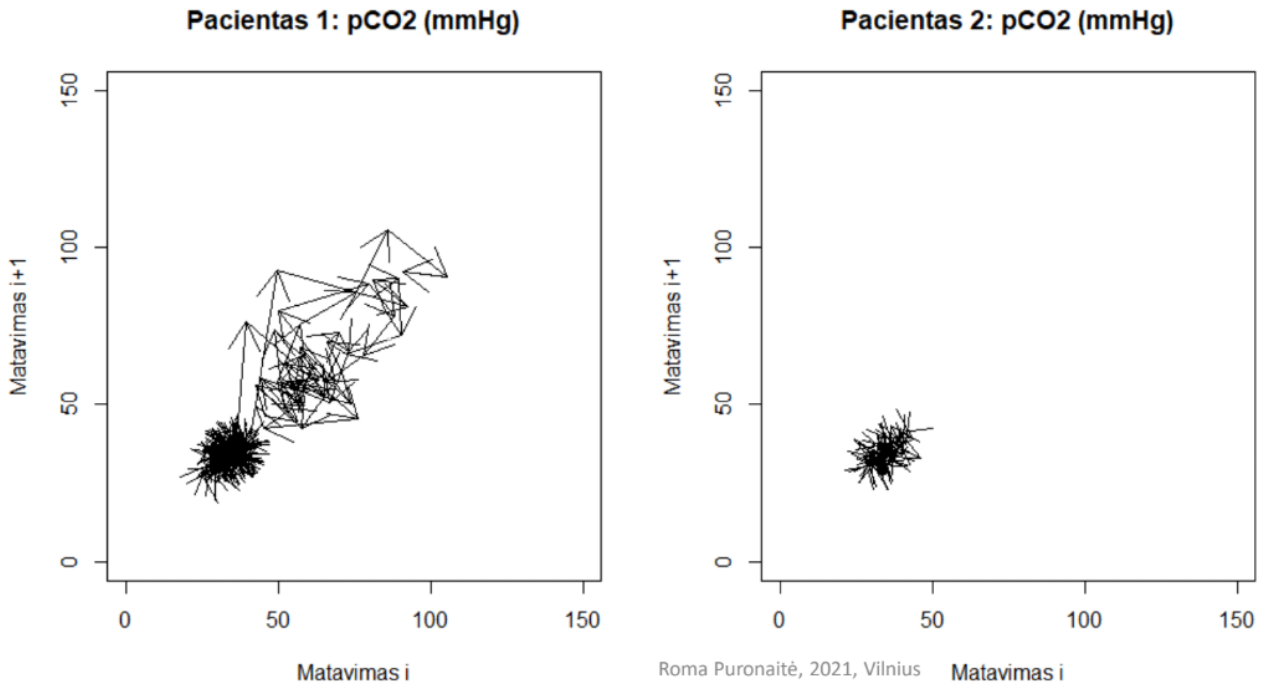
- 60-69 metų amžiaus grupė
- Vyras
- Lovadieniai: 38
- RITS
- Gydymo rezultatas: Išrašytas
- Diagnozės: Kvėpavimo nepakankamumas, Pneumonij COVID-19
- Laboratoriniai tyr.:

CRB (mg/l)	eGFR (CKD-EPI) (mL/min/ 1.73 m^2)	IL-6 Interleukinas-6 (ng/l)	K (mmol/l)	Na (mmol/l)	Šlapalas (mmol/l)
133,4	109	4,51	3,9	139	8,22

8 pav.: Eksperimentinei analizei atrinktų pacientų duomenys: demografiniai, hospitalizacijos trukmė, diagnozės ir keli laboratoriniai tyrimai atlikti pacientui atvykūs į gydymo įstaigą

Ataskaita už II mokslo metus

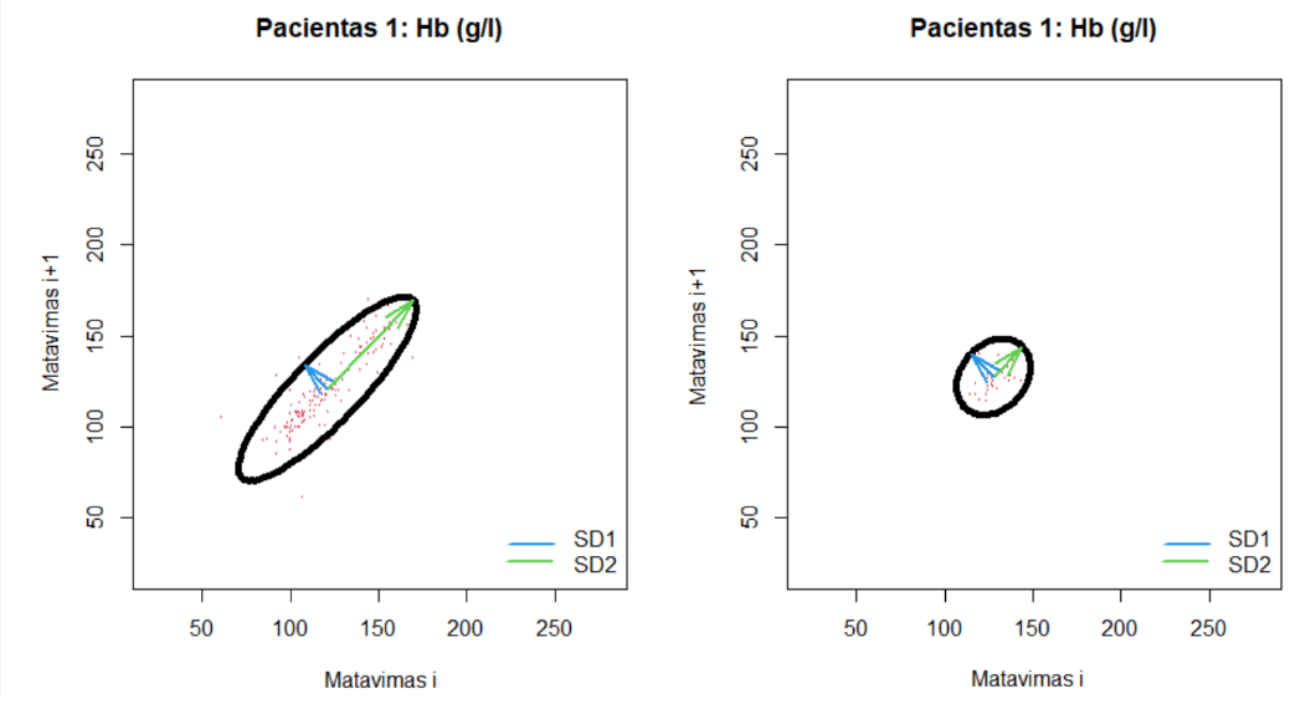
1 Eksperimentai:



9 pav.: Poincare plot: kraujo dujos (pCO2, mmHg), sujungta ir rodyklėmis atvaizduota kaitos laike trajektorija.

Ataskaita už II mokslo metus

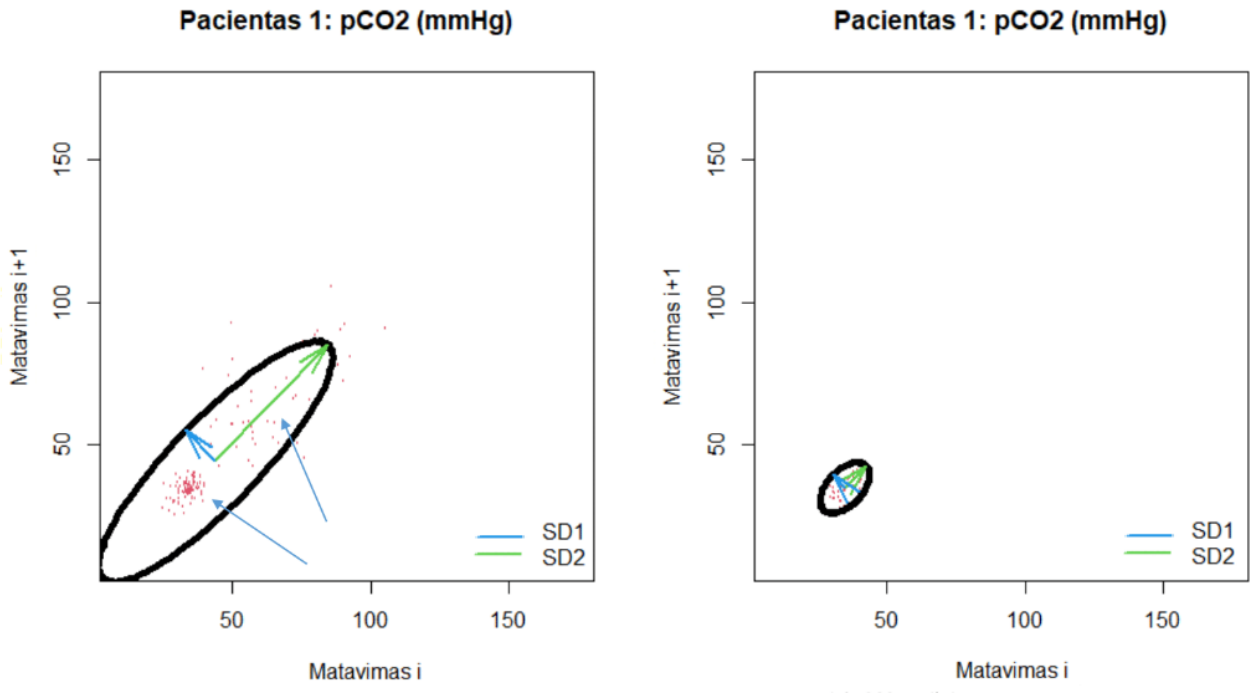
1. Eksperimentai:



10 pav.: Poincare plot: kraujo dujos (Hb, g/l).

Ataskaita už II mokslo metus

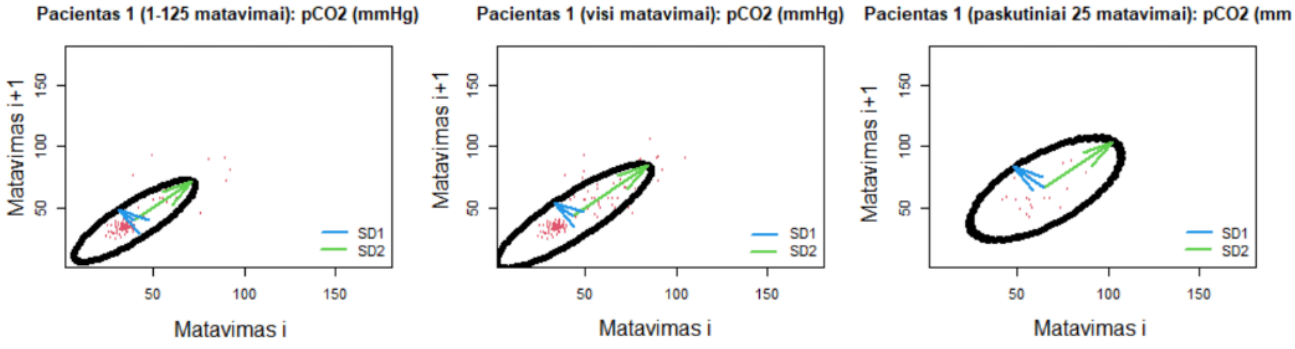
1. Eksperimentai:



11 pav.: Poincare plot: kraujo dujos (pCO2, mmHg)

Ataskaita už II mokslo metus

1. Eksperimentai:



12 pav.: Poincare plot: skirtingomis laiko atkarpomis

III (2021/2022) mokslo metų I pusmečio planas

1. Atlikta literatūros analizė parodė fenotipavimo laike (angl. temporal phenotyping) metodų taikymo poliligotumo ar COVID-19 duomenų rinkiniams trūkumą.
2. Dauguma taikomų klasterizavimo metodų neatsižvelgia į paciento būklės, ligų dinamiką laike, apsiribojama skerspjūvio tipo analize.
3. Atlikti eksperimentai rodo Poincare metodo taikymo laiko eilučių variabilumui vertinti potencialą COVID-19 EHR duomenų rinkinio kontekste.
4. Taip pat ir fizinio aktyvumo prediabeto tyrimo atveju, nustatyta reikšminga koreliacija tarp kintamumo (variabilumo) indeksų ir laboratorinių rodiklių pokyčių.
5. Kaip papildomą požymį apie būklės dinamiką (laikas tarp vizitų, laikas tarp pirmųjų diagnozių, laboratorinių tyrimų analičių verčių ir pan.), Poincare indeksus planuojama išmėginti su poliligotų pacientų duomenimis.

III (2021/2022) mokslo metų I pusmečio planas

1. Egzaminas: Daugiamačių duomenų vizualizavimo metodai (I pusm.)
2. 13th International Conference BIOMDLORE 2021 m. spalio 21-23, Vilnius, Lietuva
3. Publikacija (Technology and Health Care (IF 1.285 - statusas atspausdinta)
4. Publikacija (Darbinis pavadinimas: "Multimorbidity, depression and anxiety patterns: temporal phenotyping approach"- statusas parengta)
5. Egzaminas: Netiesiniai statistikos modeliai masinių duomenų analizėje (II pusm.)