電子医療記録を用いた Latent Dirichlet Allocation による疾患共起トピックと個別疾患ごとのコスト期待値算出

田村夏生[*1] 森雅也[*2] 上村博輝[*3] 野中尋史[*2]

*1 長岡技術科学大学 工学部 情報・経営システム工学課程 4年(B4)

*2 長岡技術科学大学大学院　工学研究科 情報・経営システム工学専攻

*3 新潟大学医学部

研究背景 > 日本，地方の高齢化

少子高齢化による医療費増大が問題

- 年間医療費は**43兆円**(2018)
- **3人に1人**が**65歳以上**の高齢者(29.1%)
- **高齢者の医療費**高騰

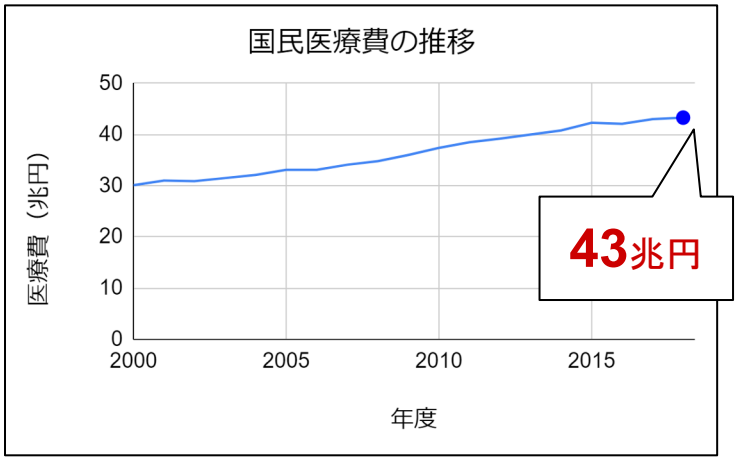
医療費の削減が重要

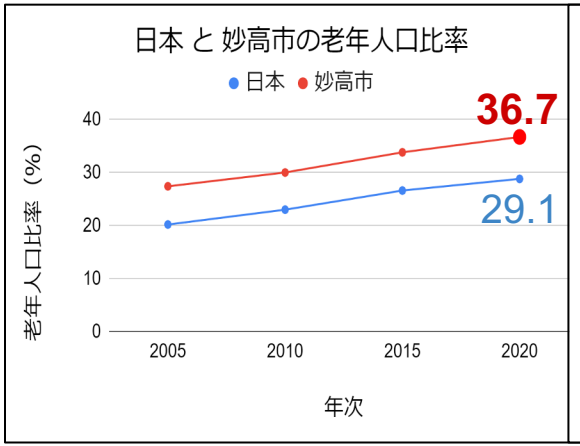
地方(新潟県妙高市)はより深刻

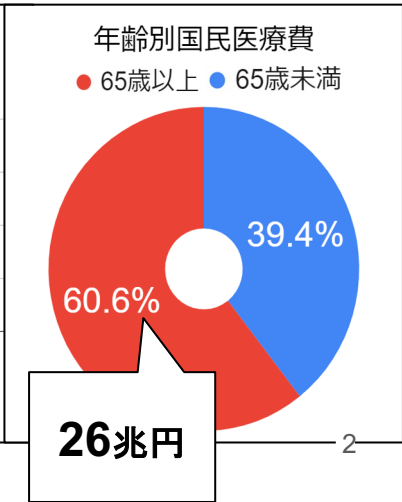
- 日本の**2040年と同じ**老年人口比率
- 老年化指数:356 **子1:老人3.5**

地方の分析が今後のモデルケースとなる

疾患・医療費予測の手法が注目







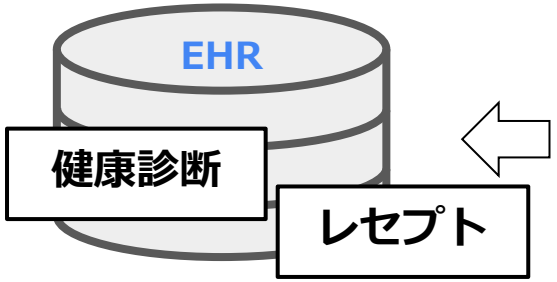
研究背景 > 疾患・医療費予測に用いられるEHR

電子医療記録が注目

電子医療記録
(Electric Health Record:**EHR**)

医療施設を超えた診療情報の記録

- 個人の**健康記録がすべて連携**
- **時系列データ**として分析可能



レセプト
病院の診療時にもらう診療行為の明細書

レセプトの問題
電子レセプトにおいて,
傷病とその診療点数（コスト）は1対１で対応していない
- **疾患のコストは処置や薬の合計**
- **疾患と処置，薬の対応わからない**

→疾患の治療費の予測が困難

病院名・患者名・疾患名・処置	
技大病院	
技大　太郎　男　65歳	
来院回数	4
うつ病	
糖尿病	
ピグアナイド薬	100
シオゾーラ5mg	50
合計点数	150

研究目的

EHRを用いた先行研究

- **レセプト→BEHRT→疾患予測(海外)**[1]
  160万人の大規模データで深層学習
  → **地方都市では大規模データの収集が困難**
- **地域健康予測**[2]
  3年以内の2型糖尿病の新規発症を予測
  → **特定の疾患のみ**
- **生活習慣病の医療費予測**[3]
  会社の従業員のレセプトで医療費予測
  → **地域のEHRでの分析が必要**
- **東京都の高齢者を分析[4]**
  高齢者の6割が3つ以上の疾患を共起
  → **共起する疾患群の特定・分析が必要**

大規模データでなくても使える

様々な疾患の共起関係を考慮可能

疾患とコスト（診療点数）の対応付け

研究目的

Latent Dirichlet Allocation（LDA）を用いた，
共起する疾患群の特定，共起する疾患群の平均コスト算出，
各疾患のコスト期待値の算出

↓

トピックベースの予測モデルの開発，現状の疾患群とコストの分析

↓

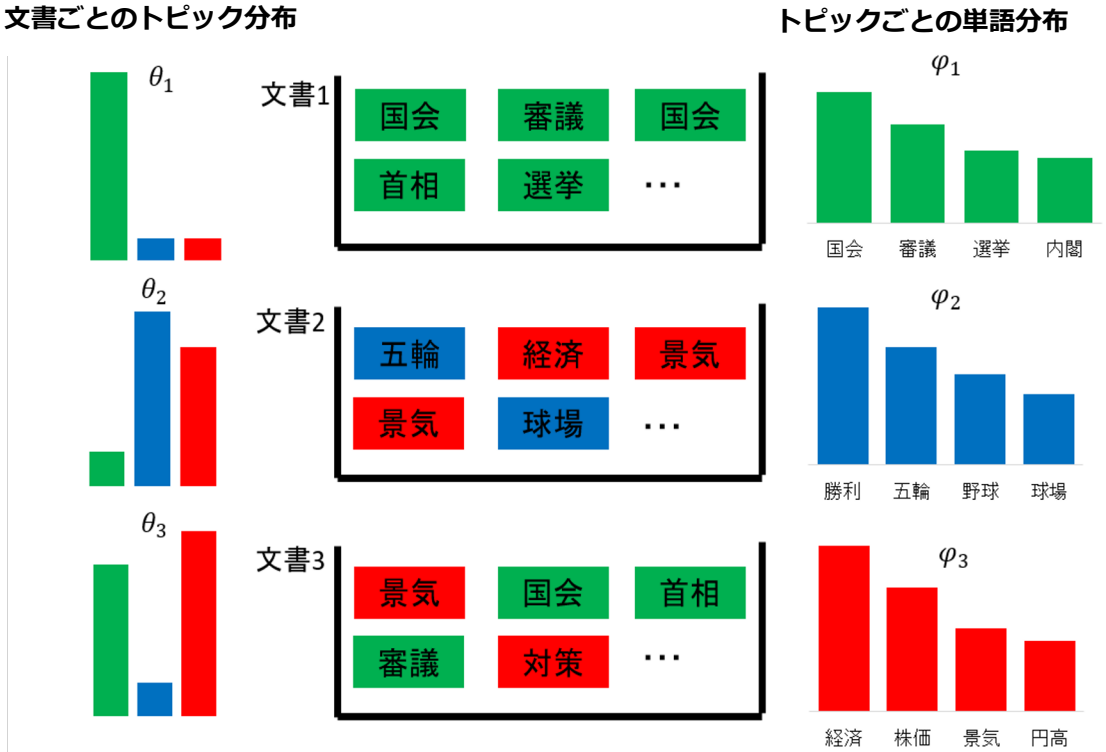
将来的に

地域のEHRで疾患・医療費予測が可能

手法 > Latent Dirichlet Allocation

潜在的ディリクレ配分法(LDA)

1つの文書に対して複数のトピックが存在すると想定した確率的トピックモデル



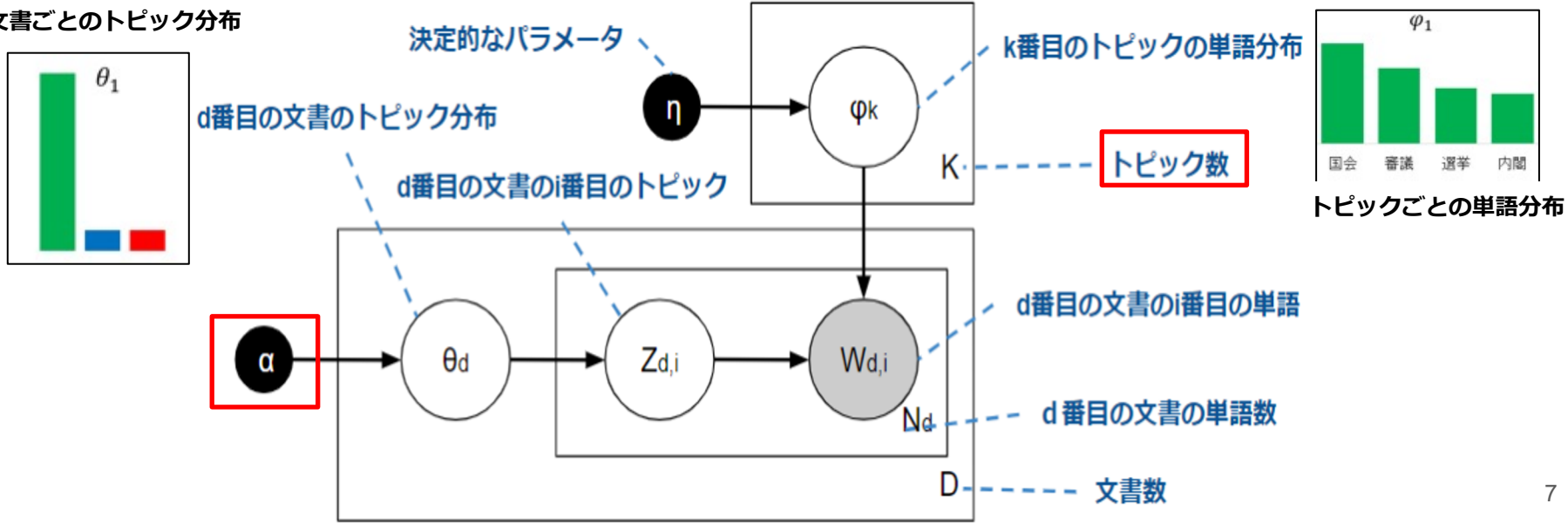
手法 > Latent Dirichlet Allocation

潜在的ディリクレ配分法(LDA)

ディリクレ分布のパラメータをα，ηとする

パラメータθ，φの事前分布としてディリクレ分布を用いる

トピックモデルのパラメータθ，φをベイズ推定する



実験データ

提供されたレセプトでは,
1か月分のある病院の患者の診療記録が
一つの文書にまとまっている

↓

同じ患者が同じ月に複数回来院していても同様

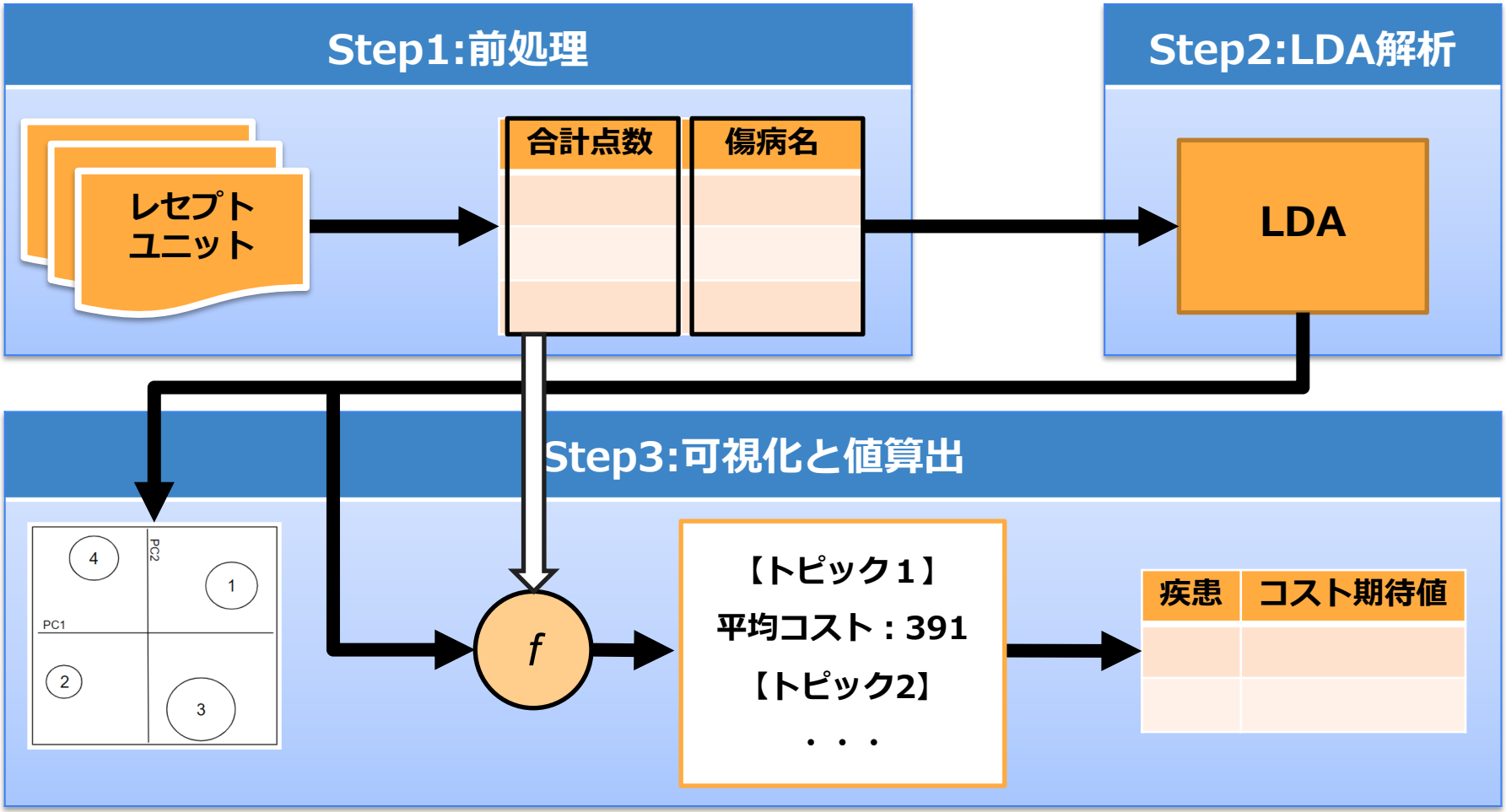
↓

定義：レセプトユニット
1か月分のある病院の患者の診療記録

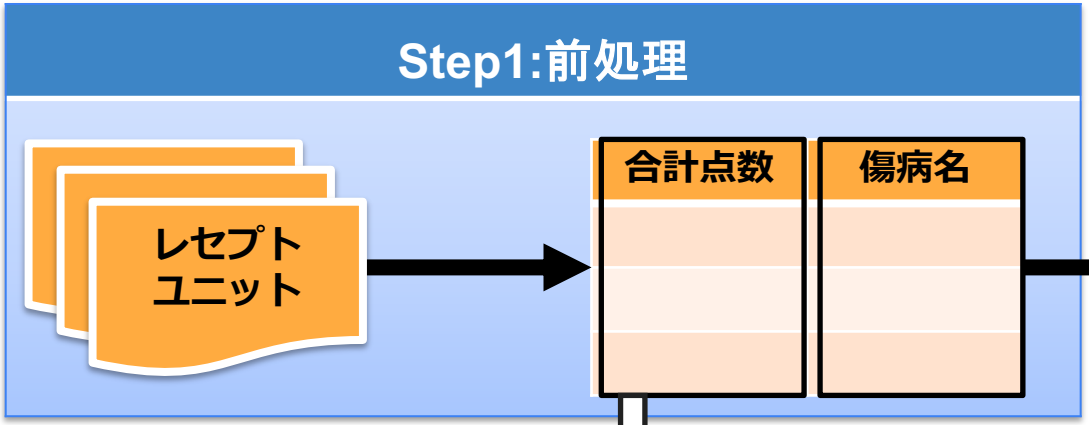
ある病院のある患者のある月のレセプトユニット

病院名・患者名・疾患名・処置	
技大病院	
技大　太郎　男　65歳	
来院回数	4
うつ病	
糖尿病	
SSRI	76
ピグアナイド薬	100
シオゾーラ5mg	50
合計点数	246

実験概要


Step1:前処理
レセプト
ユニット
合計点数
傷病名
Step2:LDA解析
LDA
Step3:可視化と値算出
PC1
PC2
1
2
3
4
f
【トピック１】
平均コスト：391
【トピック2】
・・・
疾患
コスト期待値

実験概要 ＞ Step1:前処理



平成29年3月のレセプトユニット5073件から傷病コードと合計点数を抽出

傷病コードから傷病名に変換

ストップワードを指定し，すべて除外

point	visit
282	['8835114']
4372	['4319020', '8836665', '8841622']
1237	['4660009', '8832280', '7806011', '4871001']
646	['884… '4659007']
1044	…6']

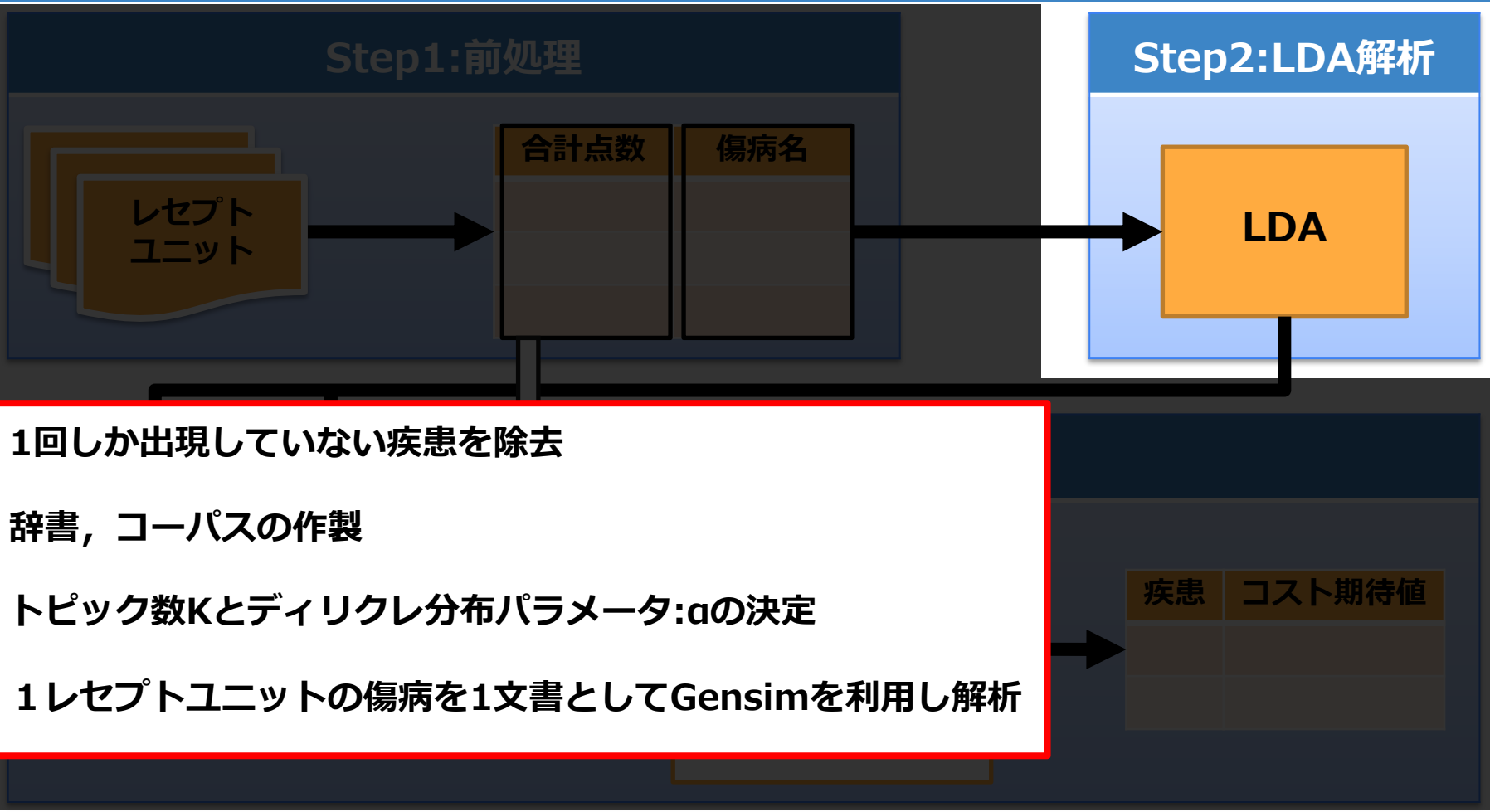
1レセプトユニットの傷病コード

	visit
	['心電図異常']
	['脳出血', '創傷感染症', '前額部裂創']
	['急性気管支炎', '急性咽頭炎', '発熱', 'インフルエンザ']
	['インフルエン… '急性上気道炎']
	['外傷後遺症… …フィー']

1レセプトユニットの傷病名

```
[['心電図異常'],
 ['脳出血', '創傷感染症', '前額部裂創'],
 ['急性気管支炎', '急性咽頭炎', '発熱', 'インフルエンザ'],
 ['インフルエンザA型', '急性上気道炎'],
 ['外傷後遺症', '関節拘縮', '反射性交感神経性ジストロフィー'],
 ['糖尿病', '慢性肝炎', '痛風', '趾関節炎', '高尿酸血症'],
 ['橈骨遠位端骨折'],
 ['関節痛',
  '筋肉痛',
  '浮腫',
  '２型糖尿病',
  '高コレステロール血症',
  '変形性腰椎症',
  '胃炎',
  'インスリン分泌異常症',
  '糖尿病性腎症'],
 ['高血圧症', '高脂血症', '高尿酸血症'],
 ['橈骨遠位端骨折', '腰痛症'],
 ['高血圧症', '心臓弁膜症', '心肥大'],
```

実験概要 ＞ Step2:LDA解析



1回しか出現していない疾患を除去

辞書，コーパスの作製

トピック数Kとディリクレ分布パラメータ:αの決定

１レセプトユニットの傷病を1文書としてGensimを利用し解析

実験概要 > Step2：LDA解析

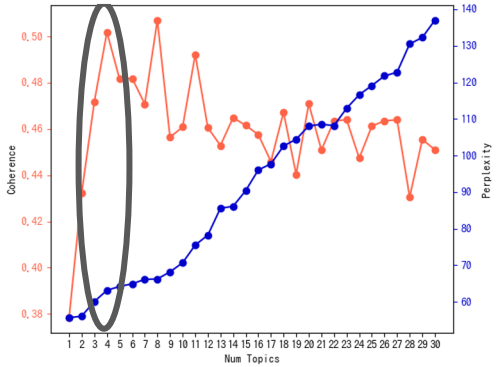
トピック数Kの決定

トピック数Kの値を変化させながら
確率モデルの精度を表す **perplexity** （PPL）と
トピックの品質を表す**coherence**(C)を算出し比較した.

トピック数＝4と決定

$$PPL = exp\left\{ -\frac{1}{N} \sum_{i=1}^{N} log_2 p(w_i|\theta) \right\}$$

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$



y軸：
- **perplexity**
- **coherence**

x軸:トピック数

ディリクレ分布パラメータαの決定

α：**トピックの選択確率を得るためのパラメータ**

αの値を変えながら**１文書あたりの平均トピック数**を算出
共起しやすい疾患が同じトピックに集まってほしいので，
1文書内の平均トピック数はより少ないことが望ましい

α＝0.01と決定

```
r = mean([len(lda[c]) for c in corpus])
```

```
alpha = 0.01, mean = 1.2856298048492016
alpha = 0.02, mean = 1.592154543662527
alpha = 0.03, mean = 1.9611669623496946
alpha = 0.04, mean = 2.314409619554504
alpha = 0.05, mean = 2.6398580721466587
alpha = 0.06, mean = 2.945791444904396
alpha = 0.07, mean = 3.1866745515474078
alpha = 0.08, mean = 3.4682239306130497
alpha = 0.09, mean = 3.502661147250148
```

実験概要 > Step3：可視化と値算出

Step1:前処理

傷病名

Step2:LDA解析

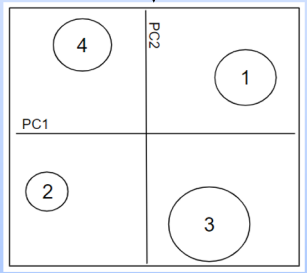
LDA

3-1. 結果の可視化

3-2. 各トピックの平均コスト算出

3-3. 疾患のコスト期待値の算出

Step3:可視化と値算出



f

【トピック1】

平均コスト:391

【トピック2】

…

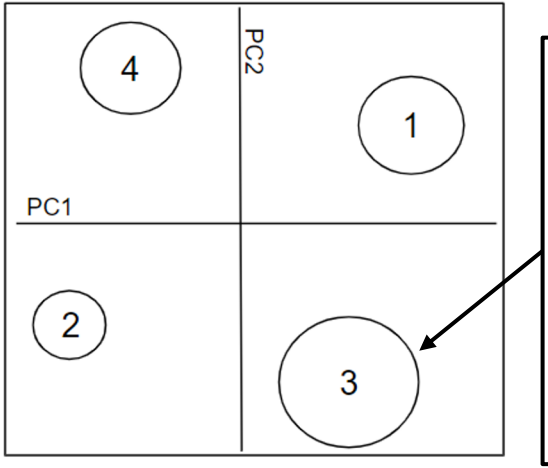
疾患	コスト期待値

実験概要 ＞ Step3-1:可視化

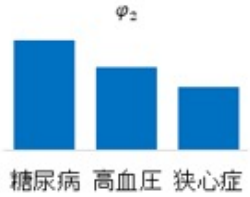
3-1 MMDSで可視化

結果を計量多次元尺度法（MMDS）で
2次元に圧縮し図に表示

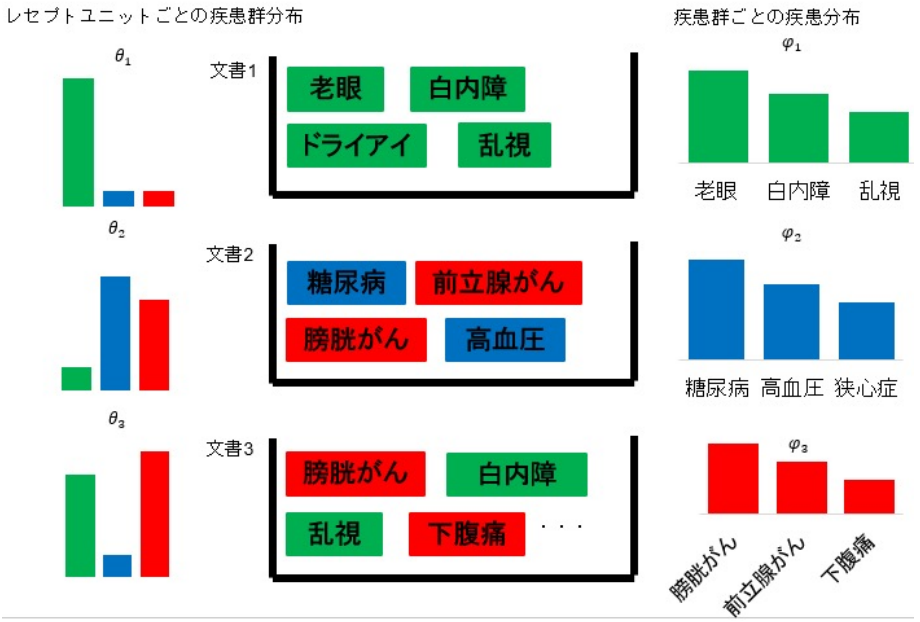
円の大きさ：トピックに含まれる文書の合計
円の距離　：トピック間の類似度



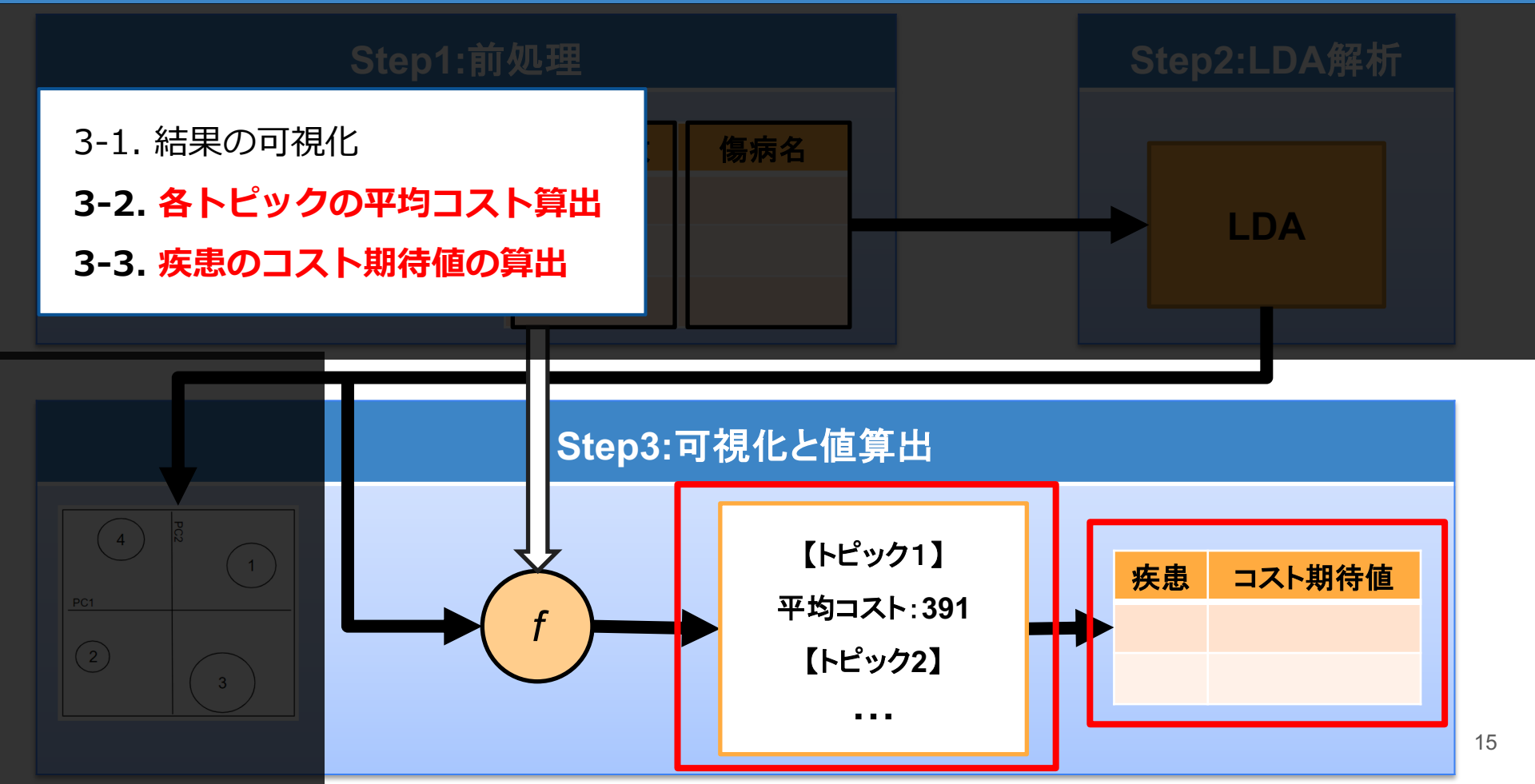
トピック：
疾患群ごとの
疾患分布



LDAによるレセプトユニット集合の生成例



実験概要 > Step3：可視化と値算出



実験概要 > Step3-2, 3-3:値算出

3-2. 疾患共起トピックの平均コスト*mj*

$$m_j = \sum_{i=0}^{N} \frac{S_i P_{ij}}{N}$$

ユニット*i*:{傷病a, 傷病b, ・・}{*Si* :100}　*Pij*:トピック*j* の割合

LDA

ユニット*i*:{ **0.1**, **0.3**, **0.5**, **0.1** }{100}　*Pij*:トピック*j* の割合

Pij × *Si*

ユニット*i*:{**10**, **30**, **50**, **10**}{100}　*N* = 5073

各トピック毎の総和/ *N*

トピック1:391, トピック2:393,
トピック3:779, トピック4:468　***mj***

3-3. 個別疾患のコスト期待値*Ejk*

$$E_{jk} = m_j P_{jk}$$

Pjk : *j*番目のトピックの*k*番目の疾患の割合

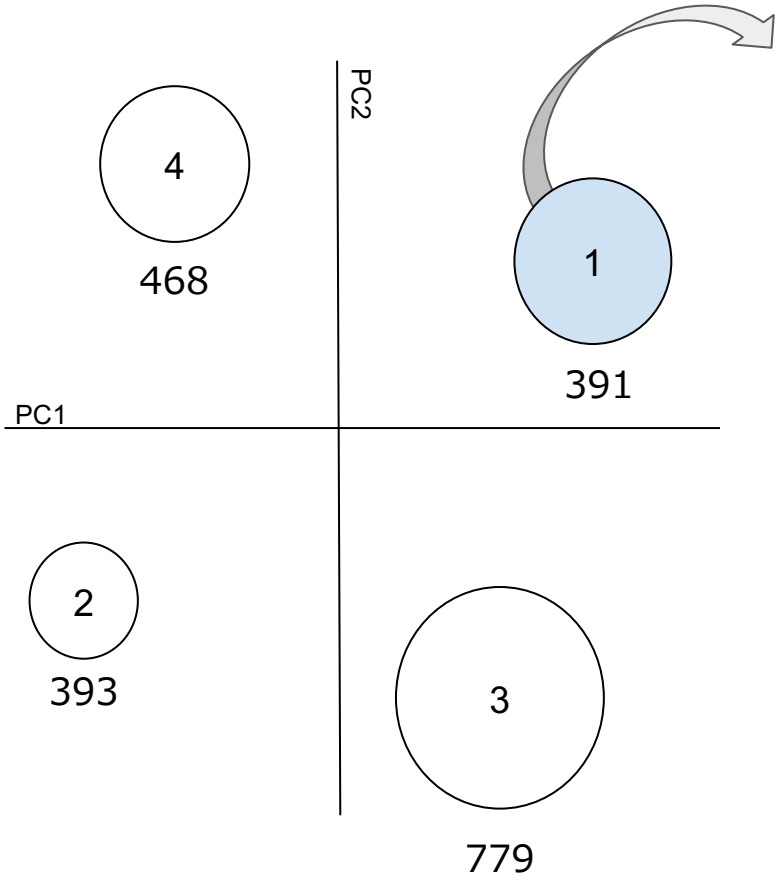
```
(0,
'0.049*"高血圧症" + 0.033*"変形性膝関節症" + 0.028*"アレルギー性鼻炎" + 0.027*"変形性腰椎症" + '
'0.027*"胃潰瘍" + 0.022*"肩関節周囲炎" + 0.021*"急性気管支炎" + 0.020*"高脂血症" + '
'0.018*"インフルエンザ" + 0.017*"閉塞性動脈硬化症"'),
```

E_{jk} = *mj* × *Pjk*

疾患共起トピックの平均コストmj

トピック*j*の*k*番目の疾患割合

結果と考察 > トピック1



結果

変形性膝関節症，変形性腰椎症，肩関節周囲炎，疼痛，腰部脊柱管狭窄，急性気管支炎，インフルエンザ，インフルエンザA型，急性上気道炎，急性咽頭炎

平均コスト：**391**

考察

特徴：**関節＋風邪**

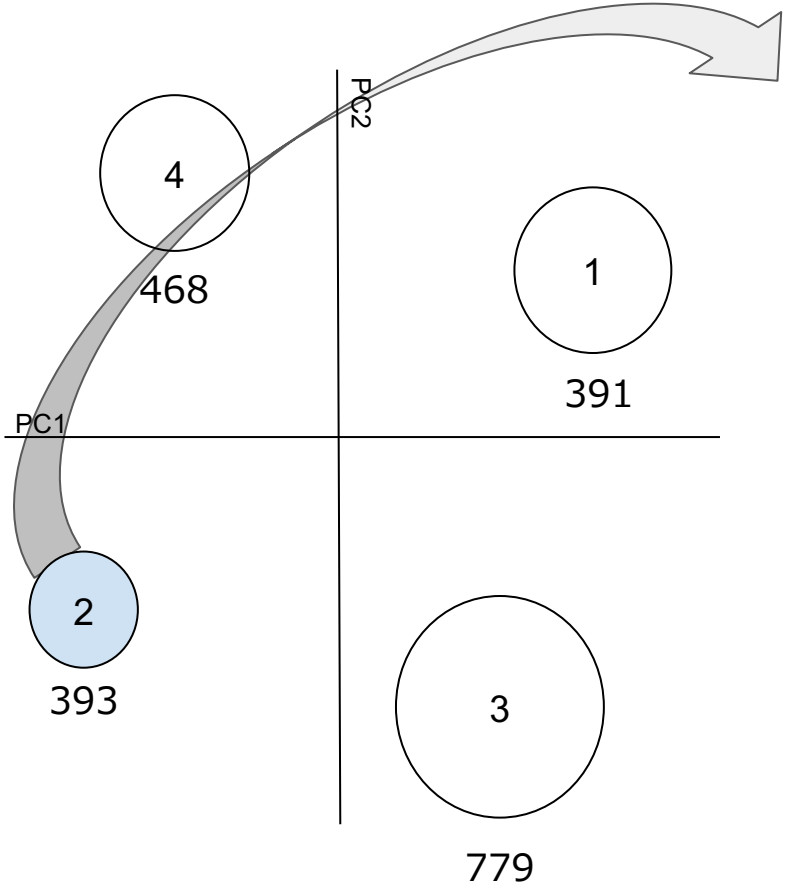
3月のレセプトのため風邪が多い

関節症と風邪が結びつくのは意外

関節症が持病で定期的に受診することで風邪が見つかる可能性

ある程度の頻度で来院→軽症のうちに発見できる可能性

結果と考察 > トピック2



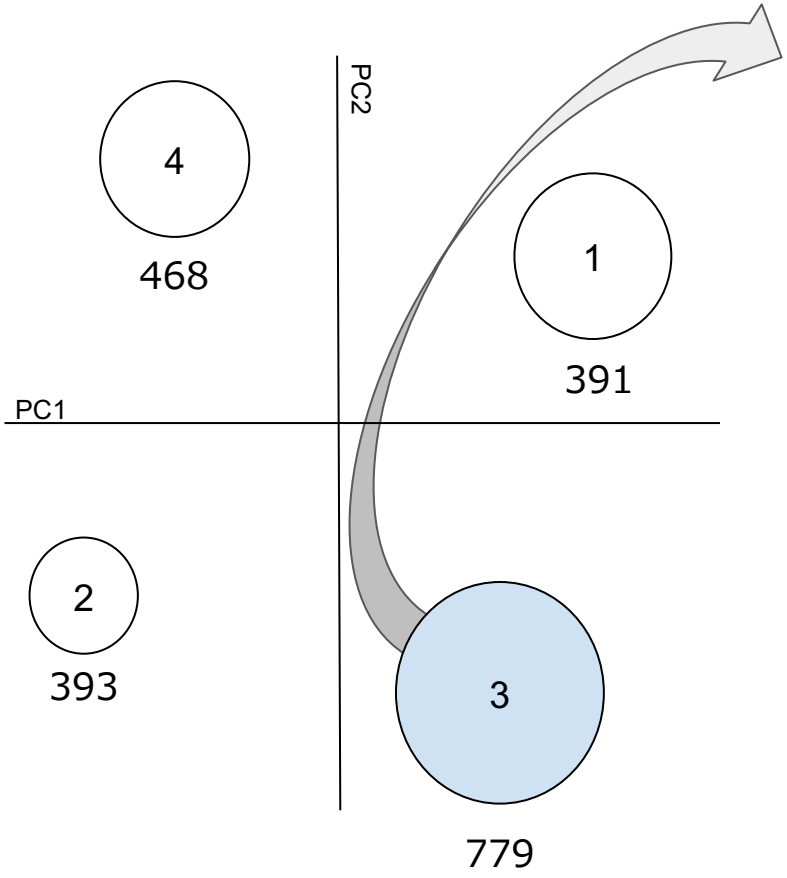
結果

前立腺がん，神経因性膀胱，過活動膀胱，下腹痛，膀胱がん，アルコール性肝炎
平均コスト：393

考察

特徴：アルコール＋男性
アルコール性肝炎と前立腺がん
→アルコール摂取量が多い男性が多い
新潟県の清酒消費量
成人1人当たり13.32L(全国一位)
新潟県の地域性

結果と考察 > トピック3



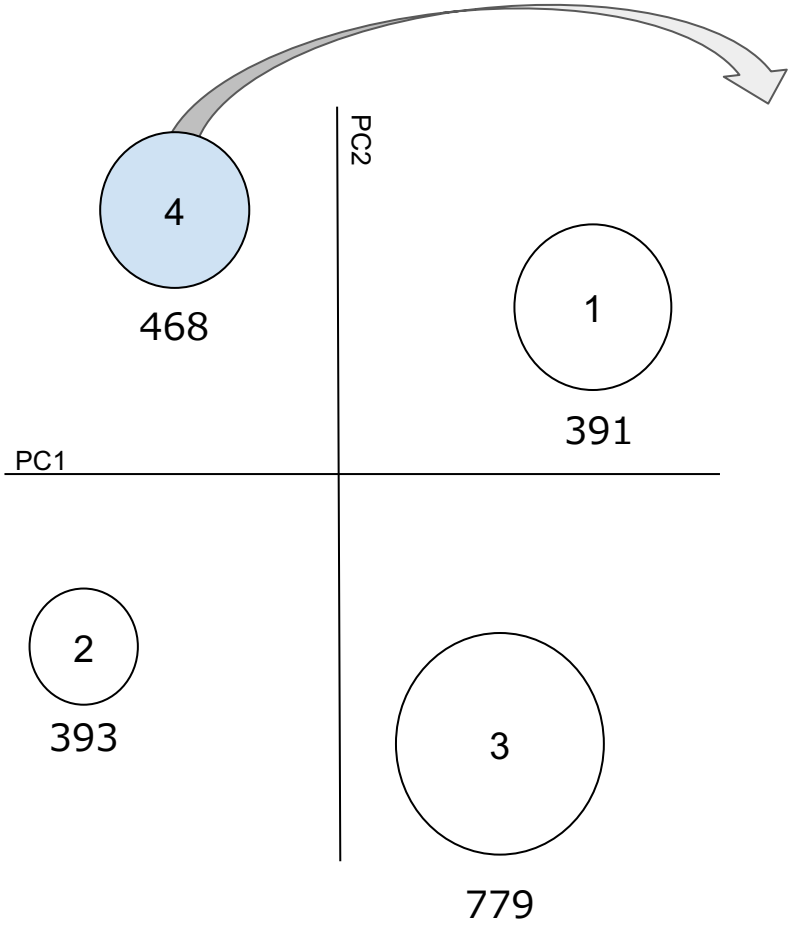
結果

高血圧症，高脂血症，不眠症，糖尿病，慢性胃炎，腎炎，脂肪肝，高コレステロール血症，高尿酸血症，狭心症，鉄欠乏性貧血，気管支喘息，逆流性食道炎，不安神経症，心肥大，閉経後骨粗鬆症，うつ病，統合失調症
平均コスト：**779**

考察

特徴：**生活習慣病**
生活習慣病が様々な疾患と共起している
合併症が多様→健康へのリスクがとても高い
治療コストが高い

結果と考察 > トピック4



結果

老視，遠視性乱視，加齢性白内障，慢性結膜炎，アレルギー性結膜炎，眼内レンズ挿入眼，近視性乱視，ドライアイ，糖尿病網膜症
平均コスト：**468**

考察

特徴：**眼＋高齢化**
加齢による眼の疾患が多く共起
モニターの見すぎなどが原因となるドライアイ，花粉などが原因のアレルギー性結膜炎など年齢によらない眼の疾患も共起

結果と考察 > 個別疾患のコスト期待値

生活習慣に関係する疾患がコスト大

高齢者に多い疾患も比較的高コスト

疾患名	点数	疾患名	点数	疾患名	点数
高血圧症	171	老視	27	慢性結膜炎	19
高脂血症	98	変形性腰椎症	26	前立腺癌	18
胃潰瘍	92	血尿	26	2型糖尿病	18
便秘症	73	遠視性乱視	24	アレルギー性結膜炎	18
糖尿病	71	加齢性白内障	22	インフルエンザ	17
アレルギー性鼻炎	56	肩関節周囲炎	21	骨粗鬆症	17
慢性胃炎	45	高コレステロール血症	21	閉塞性動脈硬化症	16
不眠症	33	前立腺肥大症	21	高尿酸血症	15
変形性膝関節症	32	急性気管支炎	20	内痔核	13

個別疾患のコスト期待値E_{jk}

$$E_{jk} = m_j P_{jk}$$

P_{jk} : j番目のトピックのk番目の疾患の割合

まとめ

妙高市の疾患・医療費分析を行うために,

共起しやすい疾患群の抽出と**疾患群のコスト**，**各疾患のコスト期待値の算出**

実験

1. 前処理
2. LDAモデルによる解析
3. 結果分析

結果

- 生活習慣病に関係する疾患群が**コスト特大**
- **地域性や高齢化の影響**が伺えるトピック
- 妙高市における**高コスト疾患の把握**

今後の展望

- これらを基礎データとして医療費の予測につなげる
- 年齢や性別を考慮した分析
- 他の地域でのデータでも分析

謝辞　データを提供してくださった妙高市に感謝申し上げます.

参考文献

[1]Li, Y., Rao, S., Solares, J.R.A. *et al.* BEHRT: Transformer for Electronic Health Records. *Sci Rep* 10, 7155 (2020). https://doi.org/10.1038/s41598-020-62922-y

[2]佐々木　健佑，久野　譜也，岡田　幸彦.ベイジアンネットワークによる地域健康予測

[3]Health Service Planning Support System Based on Medical Costs Prediction of Chronic Diseases Using Bayesian Network∗ Yasutaka HASEGAWA†a), Toshinori MIYOSHI†, Shinji TARUMI†, Takanobu OSAKI†, Hideyuki BAN†, Shouji NEGISHI††, and Norihito KUNICHIKA††

[4] Patterns of Co-Occurrence of Chronic Disease Among Older Adults in Tokyo, Japan CME ACTIVITY — Volume 16 — January 31, 2019 Seigo Mitsutake, PT, PhD1; Tatsuro Ishizaki, MD, PhD, MPH1; Chie Teramoto, RN, PHN, PhD1,2; Sayuri Shimizu, PhD3; Hideki Ito, MD, PhD4