

一般口演 | 医療データ解析

一般口演1

医療データ解析

2019年11月22日(金) 09:00 ~ 11:00 C会場 (国際会議場 2階国際会議室)

[2-C-1-01] 自由利用可能な擬似血液検査データセットの構築

○城 真範¹、興梠 貴英²、香川 璃奈³（1. 産業技術総合研究所 人間情報研究部門, 2. 自治医科大学 医療情報部, 3. 筑波大学 医学医療系 医療情報マネジメント学）

キーワード：pseudo data, laboratory test, parametric model

医学教育、医療機器等の広告など、ヒト由来データの例示が必要な場面において、個人情報が暴露されるリスクを無くし、誰もが安全に利用できる擬似的な医療データの構築を行っている。擬似医療データはまた、実際の患者から得られたデータの特徴を評価する際の基準として利用することもできる。

一般に血液検査値の分布は正規分布とはならず、値が正常であるかの判断基準は、分布を伴わない値の範囲のみで示される。このため計算機で擬似的なサンプルを出力するためには、何らかの分布を仮定することになる。本研究では、大学病院で単一の検査法から得られた実際の検査値約1万件から、Kullback Leibler divergenceを用いて、検査値の分布を4自由度を持つパラメトリックな統計分布で近似することとした。

本研究では心疾患の一部及び値が正常範囲にある擬似血液検査値のサンプルを提供する。通常、病院の患者データは健康な人のデータではないが、本研究では、疾病と無関係と思われる検査値を正常と考えることで、正常な検査値のみの抽出を実現した。具体的には、(i)各患者がもつ疾患を登録病名などの診療情報に基づいて明らかにし、(ii)疾患と検体検査結果の相関を相関ルールマイニングで明らかにした上で、(iii)(i)と(ii)の結果を利用して、各患者のうち患者に付与されている病名と相関の低い検査値を正常な検査値とみなした。検査項目の中で、特に利用頻度が高いと思われる赤血球数やHbA1cなど12種類の擬似的な正常検査値のデータセットを構築した。

検査値の分布を統計関数で表現することで数理的な検討を可能にし、高速なサンプリングを可能にした。これは、ヒト由来データの例示が必要な場面において利用できる標準的なデータセットとして応用可能である。

自由利用可能な擬似血液検査データセットの構築

城 真範^{*1}、興梠 貴英^{*2}、香川 璃奈^{*3}

^{*1} 産総研、^{*2} 自治医科大学、^{*3} 筑波大学

Development of pseudo datasets of blood test for nonprofit use

Masanori Shiro^{*1}, Takahide Kohro^{*2}, Rina Kagawa^{*3}

^{*1} AIST, ^{*2} Jichi Medical University, ^{*3} Tsukuba University

In situations where medical data needs to be exemplified, actual results of laboratory tests based on human specimen are not required; for example, medical education, advertisements for medical equipment. However, there is no published pseudo medical data that can be used in such cases. Therefore, we examined the construction of pseudo healthy humans' data from patients' data in two hospitals as a first step to create and publish pseudo medical data. First, we compared the distribution of data in the normal range from 11 blood test values. As a result, there was no significant difference in distribution among hospitals. Next, using the data of one hospital, the distribution of laboratory tests' values that is defined as normal patients' data based on the relationship between laboratory tests' values and disease names was confirmed. When the disease name was considered, the outlines of the distributions were not changed compared to the case where the disease name was not considered. However, it was found that the insufficient number of data led to the low reliability of the parameters when approximating with statistical distribution. By approximating these distributions with a lognormal distribution, a foundation was established for simply sampling pseudo blood test data. The distributions obtained by this study can be published for the situations where human-derived data needs to be exemplified.

Keywords: pseudo data, laboratory test, parametric model

1. はじめに

医学教育、医療機器等の広告など、医療データの例示が必要な場面において、必ずしも実際のヒト由来データそのものを使う必要はない場合がある。そうした場合、個人情報暴露されるリスクを無くし、誰もが自由に、安全に利用できる擬似的な医療データで代用することは有用と思われる。我々は今までに、こうした目的のため、差分方程式と乱数を利用した基本的な時系列の作成方法¹⁾や、擬似的な生存曲線などについて、大学病院の患者データをもとに作成方法を検討してきた²⁾。

様々な疾患について擬似的な医療データを整備してゆくことは有用な貢献であり、今後とも継続すべきものと考え、一方でどの疾患にも当てはまらない、すなわち健常人についての擬似データを作成することも重要と考える。それは擬似的な標準人間を仮定することであり、実際の患者から得られたデータの特徴を評価する際の基準として利用することができる。本研究では、臨床現場に於いて利用頻度が高いと思われる血液検査値について、健常人の擬似データを作成するにあたり問題となる点を整理し、病院間のデータを比較し、試験的な擬似データの分布を提案する。

2. 擬似データ作成における課題と解決

臨床現場に於いて血液検査値は、ある定められた正常範囲に値が入るかどうかにによって正常か異常か二値的に分類される。このようにすることで臨床現場における迅速な疾患の判断に寄与している。しかし擬似的な医療データを作成する観点からは、大きく二点の問題がある。まず表 1 に示すように、定められた正常範囲が、医療機関によって異なる場合がある点である(他に例えば東京大学医学部附属病院が利用している血液検査値の正常範囲は <https://www.h.u-tokyo.ac.jp/patient/depts/kensa/pdf/ketueki.pdf> から得られる)。このため、健常人の擬似データとして提供する場合に、どの医療機関の正

常値範囲を採用するかによって、端点付近の値の信頼性が低下することになる。もう一点は、一般に正常範囲内の血液検査値の分布が未知である点である。検査値が正常であるかの判断基準は、分布を伴わない値の範囲のみで示されるため、計算機で擬似的な値のサンプルを出力するためには、何らかの分布を仮定することが必要になる。

前者の課題については、もし十分に実用的な統計分布が数理的に得られていれば、データの作成者が必要な範囲を指定することで対応可能である。しかし後者については、一般に分布が左右非対称であることが想定されるため(極端な例としては正常値の範囲が片側のみで示される検査項目さえ存在する)、単純な正規分布や一様分布では適切に分布を表現することが難しいと考えられる。相当程度に真実味のある擬似データを作成するためには、何らかの実在のヒトから得られたデータを利用すべきであるが、本研究に最適と思われる健康診断のデータについては測定されるデータ種別に限りがある。

本研究では、病院で得られた患者の検査値データから、正常範囲に入っているデータ、さらには、検査項目ごとに健常人とみなせる患者を選別することで、正常範囲に入っている健常人のデータ(とみなせるデータ)を抽出し、それをもとに適切な確率密度関数を得ることとする。

3. データ

本研究では病院間の差異も確認するため、自治医科大学附属病院(以下自治医)と筑波大学附属病院(以下筑波)の患者データを利用することとした。また、検査値から何かしらの異常を見出した場合には、その検査値に"関係"する疾病名が付与されているという前提にたち、例えば「糖尿病と高血圧のみに罹患している患者は、糖尿病と高血圧に"関係"する検体検査結果以外は健常人のデータとみなせる」と考える。

検査項目(単位)	正常範囲 (自治医大)	正常範囲 (筑波大)
白血球数 (×1000/μL)	9.1 - 3.3	9 - 4
赤血球数 (×100 万/μL)	5.7 - 3.76	5 - 3.76
血小板数 (×1000/μL)	369 - 130	350 - 150
Na (mEq/L)	148 - 136	147 - 135
K (mEq/L)	5.0 - 3.6	5 - 3.6
Cl (mEq/L)	108 - 96	108 - 98
総コレステロール (mg/dL)	258 - 127	219 - 120
トリグリセリド (mg/dL)	234 - 30	149 - 30
LDL (mg/dL)	163 - 65	139 - 65
随時血糖値 (mg/dL)	109 - 73	199 - 70
HbA1c(N) (%)	6.0 - 4.9	6.2 - 4.6

表 1: 対象検査項目

そこでまず、できるだけ多くのデータを利用するため、疾病名を考慮せずに、各患者について正常範囲に入っているデータの一点(もっとも古い検査値)を、自治医と筑波にて抽出した。

さらに、筑波のデータを使い、疾病名との関連が正常値範囲内にある検査値に与える影響を考慮し、(i)各患者がもつ疾患を明らかにし、(ii)疾患と”関係”のある検体検査結果を明らかにした上で、(iii)(i)と(ii)の結果を利用して、各患者のうち健常人とみなせる検体検査結果のみ抽出するという3段階の手順をとることとした。この方法を利用することで病院データでは入りにくい健常人のデータが、検診で測定されるよりも多い項目について得られるはずである。

本研究で対象とする検査値の項目は、表 1 に示すように、両病院にて男女や年齢で異なる正常範囲が指定されていない検査項目の中で、特に利用頻度が高いと思われる赤血球数や HbA1c など 11 種類とした。材料は血液のみを対象とした。

3-1. 自治医科大学附属医院

自治医のデータについては、2013 年 1 月 1 日から 2018 年 3 月 31 までに得られた検査値データにて、疾病名を考慮せず、各患者について正常範囲に入っている最も古い一点のデータを抽出した。

3-2. 筑波大学附属医院

3-2-1. 対象患者

筑波のデータについては 2013 年 1 月 1 日から 2018 年 12 月 31 日の間に入院歴のある 20 歳以上の患者 44,516 人(平均年齢 58.5 歳、女性 52.2%)の、同期間内に採取された検査結果を対象とした。

3-2-2. 前処理

ここでは、先述した検査値と”関係”する疾病名の決定方法を説明する。

対象期間中のすべての入院歴における DPC に登録されている、入院契機傷病名、主病名、医療資源最投入傷病名、医療資源 2 番傷病名のうち疑い病名でない病名の ICD10 コード上 3 桁を患者の罹患疾患とみなした。述べ 774 種類の疾病名が認められた。

次に、患者の罹患疾患と、対象検査項目の検査値(高値と低値)との相関を、出現頻度が少ない項目同士の相関の判断に適した Lift³⁾にて評価し、検査項目ごとに上位 1 割の疾患に該当する 77 病名を、検査と関連が強い疾病名と定義した。

3-3. 統計モデル

Box-Muller 法による簡便なサンプリングと数理的評価を可能とするため、ヒストグラムは式 1 に示す対数正規分布で近似することとし、その 4 パラメータを最小二乗法で決定した。

$$y = A \exp \left(- \left(\frac{\log(x/x_0)}{\sigma} \right)^2 \right) + K_0 \tag{式 1}$$

4. 結果および評価

4-1. 疾病名を考慮せず抽出した結果

典型的なヒストグラムの比較を図 1 に示す。左側(青)が自治医で右側(赤)が筑波である。サンプルサイズは表 2 にまとめる。黒線は対数正規分布による近似であり、そのパラメータは表 3,4 にまとめる。

図1 疾病名を考慮せず抽出した場合のヒストグラム

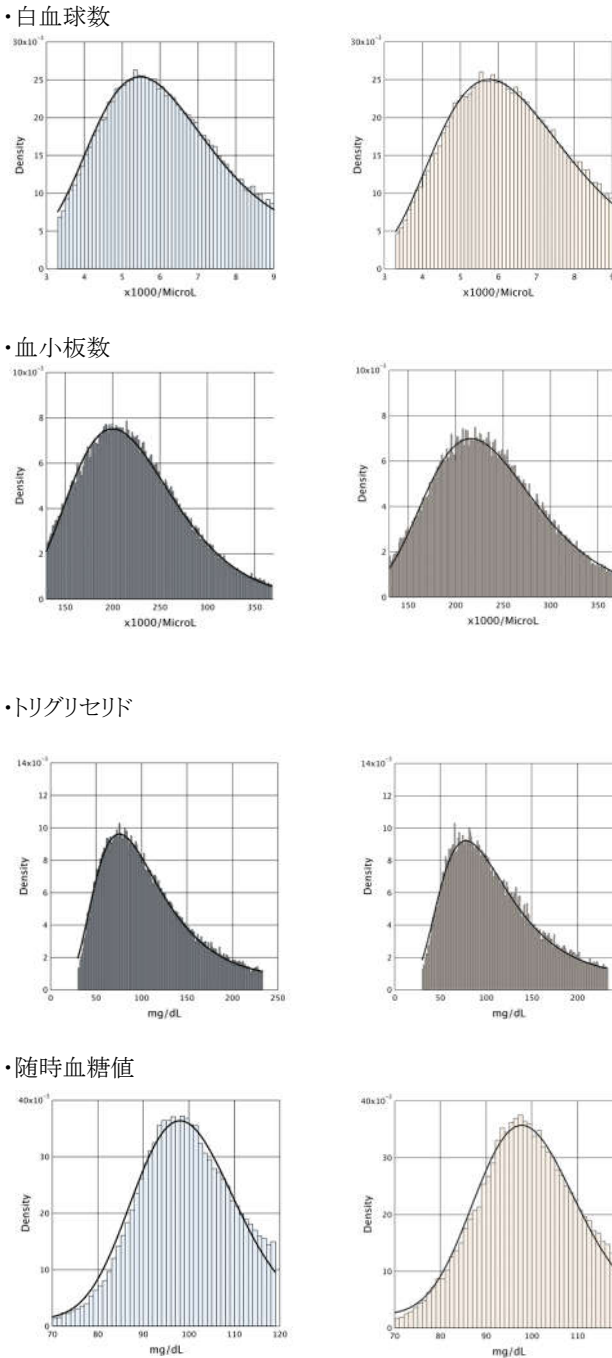


表 2 疾病名を考慮せず抽出した場合のデータサイズ

	自治医(人)	筑波(人)
白血球数	122064	68532
赤血球数	122916	73228
血小板数	134978	76414
Na	126986	79274
K	121728	75608
Cl	120108	76587
総コレステロール	65860	46998
トリグリセリド	71652	46122
LDL	46477	23429
随時血糖	78609	56108
HbA1c	46853	39613

表 3 疾病名を考慮せず抽出した分布の対数正規分布近似 (自治医)

	K_0	A	X_0	σ
白血球数	0.003467	0.021935	5.4765	0.39134
赤血球数	-0.00026	0.009436	4.3131	0.16112
血小板数	-3.08E-05	0.007541	199.96	0.38423
Na	0.013428	0.16896	141.46	0.019845
K	0.016495	0.10617	4.1651	0.099647
Cl	0.01243	0.17996	105.47	0.031657
総コレステロール	0.001717	0.009803	184.98	0.25161
トリグリセリド	0.000572	0.009038	75.525	0.67633
LDL	0.005186	0.011465	101.85	0.35251
随時血糖	0.001231	0.035145	98.056	0.16125
HbA1c	0.006046	0.10991	5.603	0.093803

表 4 疾病名を考慮せず抽出した分布の対数正規分布近似

	K_0	A	X_0	σ
白血球数	0.000663	0.024368	5.7569	0.42317
赤血球数	-6.8E-05	0.008637	4.4426	0.15971
血小板数	0.000221	0.006767	216.11	0.37261
Na	0.006296	0.17701	141.02	0.020982
K	0.0133	0.11268	4.1185	0.1008
Cl	0.010585	0.1827	104.9	0.029189
総コレステロール	-8.8E-05	0.0116	181.85	0.29269
トリグリセリド	0.000682	0.008539	77.861	0.67922
LDL	0.00466	0.012343	103.43	0.34473
随時血糖	0.00232	0.033407	97.785	0.15948
HbA1c	0.00578	0.10819	5.5597	0.095592

(筑波)

4-2. 疾病名を考慮して抽出した結果

典型的なヒストグラムの比較を図 2 に示す。左側(緑)が疾病名を考慮した結果で右側(赤)が考慮しない結果である。右側は 4-1 の結果と同じである。サンプルサイズは表 5 にまとめる。黒線は対数正規分布による近似であり、疾病名を考慮した場合のパラメータを表 6 にまとめる。

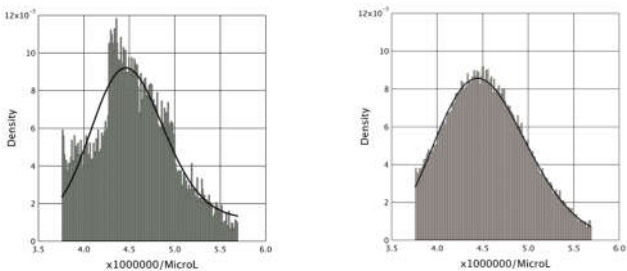
表 5 疾病名を考慮して抽出した場合のデータサイズ

	筑波(人)
白血球数	27075
赤血球数	13017
血小板数	26488
Na	25055
K	22827
Cl	25415
総コレステロール	6428
トリグリセリド	13393
LDL	3167
随時血糖	17681
HbA1c	17373

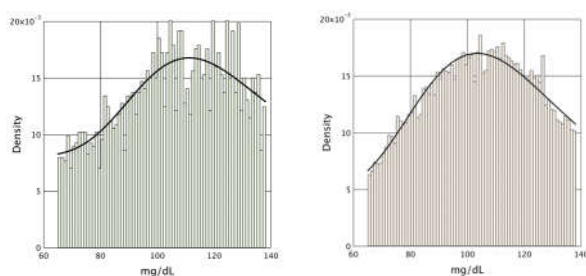
表 6 疾病名を考慮して抽出した分布の対数正規分布近似 (筑波)

	K_0	A	X_0	σ
白血球数	0.0084222	0.017525	5.6923	0.35372
赤血球数	0.0011725	0.008038	4.463	0.12391
血小板数	-0.0001045	0.0074401	200.25	0.45079
Na	0.014562	0.1681	141.05	0.020096
K	0.011691	0.1192	4.0801	0.099026
Cl	0.018214	0.17572	105.03	0.02865
総コレステロール	0.0031502	0.011422	190.59	0.2543
トリグリセリド	-0.0026562	0.013745	81.46	0.87592
LDL	0.0080502	0.008736	111.21	0.28288
随時血糖	0.0026269	0.032836	97.437	0.15951
HbA1c	0.0084197	0.101	5.5667	0.10048

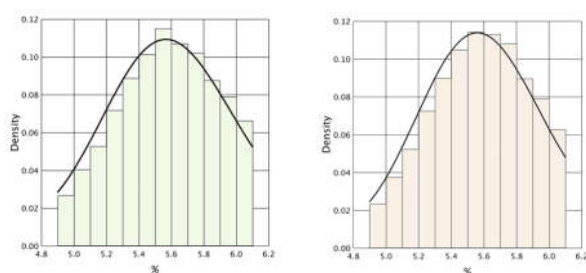
図 2 疾病名を考慮して抽出した場合のヒストグラム・赤血球数



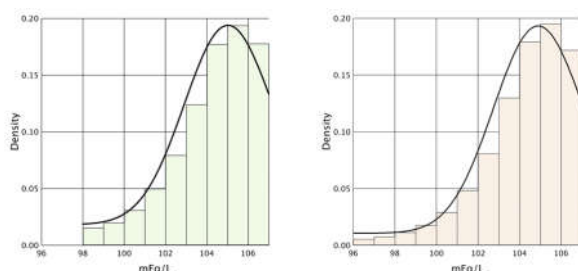
・LDL



・HbA1c



・Cl



5. 考察

疾病をもつ患者の血液検査値から正常範囲を取り出した結果、厳密には疾病の割合が異なるはずの 2 病院にて得られた分布が、ほぼ似た形となることが示された。これは正常値とされる血液検査値が疾病からの影響を受けにくい可能性、あるいは 2 病院で患者の疾病の割合が似ている可能性も考えられる。前者について確認するため、同一病院の同一データから、疾病との関係性が遠いと思われる血液検査値だけを抜いて調べたところ、やはり同様の傾向の分布を得た。後者については今後、検診データなどとの比較、確認をする必要があると考えるが、本研究で調べた範囲においては、病院が保有する血液検査値についても正常値とみなされる範囲の検査値を抽出することで安定した正常値範囲の分布を得られることが示された。

擬似的な血液検査値を簡便に得るために、本研究で利用したヒストグラムから統計密度関数を構成した。分布による近似においては、対数正規分布による他、ジョンソンの SU 分布や 2 個の正規分布の和(ガウシアン展開)も試したが、本研究で扱ったような平時より安定値が存在する血液検査値においては対数正規分布によるものが一番適切であった。異常時にだけ大きな値を取る血液検査値について、どのような分布が適切であるかは今後の課題である。また、なぜ対数正規分布による近似が適切であるかについても今後の検討を必要とする。

今回は 3 種のパラメータセットを並列的に提示している。ただし疾患との関係を考慮した場合では、もっとも顕著な場合、母数が 10% 程度に減るため、分布のばらつきが大きくなり、パラメータの信頼性は低下するものと思われる。

6. まとめと展望

本研究では病院の保有する患者データから健康な人のデータを構成することを検討した。

まず 2 つの病院の保有する 11 種の患者血液検査値から、疾病名を考慮せず正常範囲にあるデータを抽出、比較した。その結果、病院間の分布には大きな差が見られなかった。次に筑波大学附属病院のデータにて、疾病名を考慮してより正常範囲にあると思われるデータを抽出し、疾病名を考慮しない場合と比較した。その結果、疾病名を考慮する場合、分布の概形に大きな変化は見られなかったが、条件を厳しくすることでデータ点数が不足し、統計分布にて近似する際に、パラメータの信頼性が下がることがわかった。

これらの分布を対数正規分布で近似することで擬似的な血液検査データを簡便にサンプリングするための基盤を整備した。本研究によって得られた分布は、ヒト由来データの例示が必要な場面において、自由かつ無償で利用できる。

今回の結果は 2 病院のみの結果であるので、今後より多くの病院データにて確認するべきである。また、病院でしか得ることのできない検査値についても分布を明らかにし、将来的には疾病の程度を定量化するための標準として展開してゆきたい。

7. 倫理的配慮

本研究は自治医科大学倫理審査承認番号臨大 17-140、筑波大学附属病院研究倫理審査委員会(承認番号 R01-080)の承認を得ている。適切な検査値の抽出とヒストグラム化は全て、倫理審査を経て許諾された範囲にて、興梠と香川が行った。城は検査値データの閲覧、操作をしていない。

8. 謝辞

本研究は産業技術総合研究所イノベーション推進本部経費「健康人血液データの標準化」、および科研費 19K19347 の支援を受けている。

筑波大学附属病院のデータ抽出と管理については筑波大学医学医療系の讃岐勝氏にご尽力頂いた。本研究の標準化の側面については産総研物質計測標準研究部門バイオメディカル標準研究グループの藤井紳一郎氏、産総研人間情報研究部門の赤穂昭太郎氏に助言を頂いた。本研究において他に開示すべき COI はない。

参考文献

- 1) 城 真範, 興梠 貴英, 医学用擬似生体データの生成, 第35回医療情報学連合大会 プロシーディングス.
- 2) 城 真範, 興梠 貴英, 赤穂 昭太郎, 山崎 力, 永井 良三, 生存時間解析用擬似データ作成システム, 第 36 回医療情報学連合大会 プロシーディングス.
- 3) Tan, P. N., Kumar, V. and Srivastava, J.: Selecting the right objective measure for association analysis, Information Systems 2004; 29: 293-313.