Vol.2014-DBS-160 No.27
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.27
2014/11/18

創薬データマイニングにおける副作用予測モデルの提案

江谷典子[†1][†2]

近年の新薬開発では，既存薬の作用から新たに薬効を見つけ出し，別の疾患に対する治療薬として再開発する創薬研究が注目を浴びている．そこで，大規模なデータの集合体に対しての手法の適用によるモデルを作成することで，何らかの新事実・関係を発見するという立場から，ビッグデータを応用した創薬（Drug Discovery）データマイニングの研究に取り組んでいる．本研究では、化合物とタンパク質の相互作用データベース STITCH4.0 に収録されているデータと副作用データベース SIDER 2 に収録されているデータから薬と副作用の関係を網羅的に予測できるモデルを構築し提案する．PLS 回帰式モデルを用いた判別分式を導入することで，副作用発症率分類の特徴抽出を可能にし，2 値化することができる．この 2 値化により，化合物とタンパク質の組み合わせにおける副作用発症率を「41%～100%」「0.1%～40%」の 2 段階に分けることを示す．さらに，サポートベクターマシン（SVM）を用いて，副作用の発症率をこの 2 段階で予測することができることを示す．

Side Effect Prediction Model in Data Mining for Drug Discovery

NORIKO ETANI[†1][†2]

In the recent new drug development, the research on drug discovery, that its indication is newly found out from the approved drugs and a new drug is redeveloped with the new indication as a new therapeutic drug for a different disease, has attracted attention. I have researched and developed data mining for drug discovery as a big data application from the standpoint that a model is developed by the methods for collection of large-scale data in order to discover some new facts and relationships between data. In this paper, a model that can predict comprehensively the relationship of side effects and drugs will be proposed from the data on the side effects database "SIDER 2" and the data on the chemical-protein interaction database "STITCH4.0". This paper describes that the feature of side effect incidence is extracted and presented as two categories by introducing discriminant analysis using PLS regression model, and that the chemical-protein interaction is classified into two stages of "41%-100%" and "0.1%-40%". Moreover, it describes that support vector machine (SVM) can predict side effect incidence by these two stages.

1. はじめに

近年の新薬開発では，既存薬の作用から新たに薬効を見つけ出し，別の疾患に対する治療薬として再開発する創薬研究が注目を浴びている [1][2]．その理由の一つは，”one-drug-one-gene”アプローチでは十分に治療ができないほど，人間の疾患は複雑な生物学的プロセスであることにある [3]．そこで，人間での安全性と体内動態が臨床で充分に確認されている既存薬の作用を徹底的に調べ，それを活かして，その既存薬を他の疾患治療薬として開発する研究が求められている [4]．新たな薬効を示す医薬品は，目的とする効能以外ではむしろ副作用を示すことになる問題点がある．処方薬の副作用に関する公式報告は，この１０年間で劇的に増加しているという．米食品医薬品局（FDA）は，2011年現在，そのような医療品に関連する健康被害，ときには死亡の報告を，毎年約50万件受け取っている [5]．薬剤の安全性を考える上で，副作用を創薬段階で予測し，回避することは重要な課題である．一方，新薬の創薬手法は，有機合成などによって得られた物質をスクリーニングする手法から，遺伝子情報を基に標的分子を見定め，それに作用する化合物をスクリーニングで見つけるゲノム創薬に大きく転換している．従来の創薬プロセスは，数万存在する化合物を一つずつ調査することをベースにしていたため，新しい医薬品の候補物質の発見には研究者の経験だけでなく，勘や偶然といった面に頼らざるを得ない面があった．しかし，遺伝子情報をベースに病気との関連性を解析し，論理的かつ科学的に新たな医薬品の可能性を見つけようとするゲノム創薬は，コンピュータに蓄積されたデータを用いて検証するだけなので，効率が非常によいのが特徴である [6]．現在，オーダーメイド医療の取り組みにより，病気のかかりやすさ，薬剤への感受性，副作用を予測できる SNP（スニップ）データを用いて，新薬や治療法を開発していくことや遺伝子の特徴が合うか合わないかを調べて副作用を回避することが期待されている [7]．そこで，新たな薬効の発見においても，副作用を避けることができるように，薬候補を発見できるような創薬（Drug Discovery）データマイニングを目指して，研究開発に取り組んでいる．

データマイニングとは，一般的には「大規模なデータの集まりから，価値があり自明でない情報を効率的に発見することである」という定義付けがなされている [8][9]．すなわち，知識発見という考え方がその主流となっているものである．重要な点は，データマイニングとは，統計的な仮説検定やすでに設定された仮説についての検証ではなくて，「データ間の新事実や関係の発見」ということである．事前に設定された仮説があり，それを検証するという方法

†1 京都大学
Kyoto University
†2 独立行政法人科学技術振興機構, CREST
CREST, Japan Science and Technology

Vol.2014-DBS-160 No.27
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.27
2014/11/18

ではなく，データの集合体に対して，統計や人工知能の手法を適用するためにモデルを作成することで，そのなかに何らかの新事実・関係を発見するという立場に立つものである．本稿では，このデータマイニングの考え方を基に，ビッグデータを応用した創薬（Drug Discovery）の研究に取り組む中で，まず，現在報告されている副作用に関する臨床データを用いて，創薬データマイニングにおける副作用を回避できるように副作用予測モデルを構築し，提案する．

本論文の構成を以下に述べる．まず2節で関連研究について述べる．次に3節で本提案における副作用予測モデルの構築を説明した後，4節でその有効性を確認するための評価を示す．最後に5節でまとめる．

2. 関連研究

化合物とターゲットタンパク質との関係から薬の副作用を予測する試みが行われている．例えば，Kuhn ら [10]の研究が挙げられる．薬の副作用における類似点に着目して，「タンパク質と副作用」の組み合わせを証明するために「化合物とタンパク質」に関する臨床試験のデータを集積し，その結果，1428件の副作用のうち，個々のタンパク質から732件を予測し，その732件のうち137件は薬理学的データが存在することで証明できるとしている．さらに，大規模な副作用の予測を立てるには，「薬とタンパク質」および「タンパク質と副作用」の関係の計算上のモデルとして捉えるような手法が必要である．

計算上のモデルとして捉えるような手法については，Yamanishi ら [11]がカーネル回帰モデルを提案している．化合物の構造とターゲットタンパク質の情報から潜在的な副作用を予測する手法である．この手法により，969件の認可されている薬の副作用の予測ができた．薬の副作用プロファイルでは，各薬は，969（副作用数）次元の特徴ベクトルで表現される．化合物プロファイルでは，各薬は，881（フィンガープリントにおける化合物の部分構造数）次元の特徴ベクトルで表現される．生物学的プロファイルでは，化合物とタンパク質の相互作用の特徴を示すために，各薬は，1368（化合物と結合できるタンパク質の数）次元の特徴ベクトルで表現される．これらの関係をカーネル回帰式により求める．

これらの研究は，大規模なデータを逐次的に処理を行いながら副作用の予測を行っていると言える．また，薬の特徴として「化合物とタンパク質との相互作用」に着目するといったアプローチは創薬の観点からも明らかな関係であるものの，薬の薬理的作用との関係が明らかではない．

本研究では，データマイニングの手法を取り込んで，薬の副作用と係る最小限のデータモデルから予測モデルを構築し，大規模なデータにおけるルールとなるような工学的アプローチにより，前章で述べたような問題の解決に取り組んでいる．

3. 副作用予測モデルの構築

3.1 薬の特徴

薬の特徴には，生物学的特徴として化合物と結合できるタンパク質，臨床的特徴として効能・ATC コード・副作用がある [3]．ATC コードとは，医薬品の分類に用いられる解剖治療化学分類法で，WHO の医薬品統計法共同研究センターによって管理されている [12][13][14]．この分類法では，医薬品は効果をもたらす部位・器官，および作用能・化学的特徴によって５つのグループに分けられる．以下の 5 つのレベルによる分類法が適用される（表 1）．

表 1　ATC コードの構成[13][14]

第 1 レベル：解剖学的部位に基づいた分類で，アルファベット 1 文字で表される．
A　消化管および代謝
B　血液，および血液を生成する器官
C　循環器系
D　皮膚
G　泌尿生殖器系，性ホルモン
H　全身性のホルモン調節剤，性ホルモンとインスリンを除く
J　全身性の抗感染症薬
L　抗悪性腫瘍薬，免疫調節剤
M　筋骨格系
N　神経系
P　駆虫性薬剤，殺虫剤，忌避剤
R　呼吸器系
S　感覚器系
V　その他（診断薬，一般栄養剤など）
第 2 レベル：治療法メイングループによる分類．2 個の数字で示される．
第 3 レベル：治療法・薬学サブグループによる分類．1 個のアルファベットで示される．
第 4 レベル：化学・治療法・薬学サブグループによる分類．1 個のアルファベットで示される．
第 5 レベル：化学構造サブグループによる分類．2 個の数字で示される．

例えば，ミトキサントロン（Mitoxantrone）には，ATC コード「L01DB07」が付与されている．このコードは，次のように分類されている．

第1レベル「L」 抗悪性腫瘍薬，免疫調節剤
第2レベル「L01」 抗悪性腫瘍薬
第3レベル「L01D」 細胞障害性抗生物質と関連物質
第4レベル「L01DB」 アントラサイクリンと関連物質
第5レベル「L01DB07」 ミトキサントロン

Vol.2014-DBS-160 No.27
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.27
2014/11/18

3.2 データベース

副作用予測モデルでは，薬の特徴より「ターゲットタンパク質との相互作用スコア」「ATC コード第1レベル」「発症する副作用」「副作用発症率」を用いる（図 1）．対象とする化合物は，データベース SIDER 2に収録されている薬の内，0.1%以上の副作用を発症している化合物335件とした．SIDER 2とは，欧州19か国の出資により1974年に創設された欧州分子生物学研究所（European Molecular Biology Laboratory ： EMBL）により副作用リソースの一部を CCO として用いて，パブリックドメインとし，制限なくアクセスできるようにしたデータベースである [15]. SIDER には，市場に流通している薬物や，これらの記録された不都合な副作用，薬物反応に関する情報を掲載している．現在，996の薬および4192の副作用が収録されている．また，ATC コード・発症する副作用・副作用発症率も SIDER 2から抽出を行った．さらに，対象となる化合物と結合できるタンパク質は，データベース STITCH 4.0に収録されているものを抽出し，その相互作用スコアを用いる．STITCH 4.0とは，前述の EMBL により，390,000の化合物と36万のタンパク質の相互作用ネットワークを収録した各種 DB ならびに文献をキュレートしたデータベースである [16][17]. 化合物とタンパク質の結合度を示す相互作用スコアが提供されている．対象となる化合物335件（2014年10月1日現在）の各結合度上位10位を抽出した．

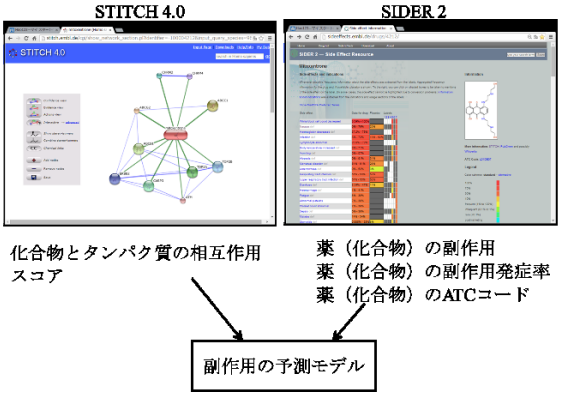

図 1　利用するデータベースとデータ

3.3 データ分析

(1) 発症する副作用

発症する副作用を中心に，357件のデータモデル「ターゲットタンパク質との相互作用スコア上位第1位」「ATC コード第1レベル」「発症する副作用」「副作用発症率」を分析した．その結果，解剖学的部位・器官を示す「ATC コード第1レベル」に基づき分類することができた（表 2）．そこで，本稿では，副作用の発症率について予測ができるモデルを構築することにした．発症する副作用の予測については，今後，遺伝子情報との関連を分析する際に検討する．

表 2　ATC 第 1 レベル別発症分類

A	消化管および代謝
	血糖，下痢，感染，嘔吐，疾患の進行
B	血液，および血液を生成する器官
	貧血，出血，膿瘍，体液過剰
C	循環器系
	血腫，静脈炎，頻脈，浮腫，肝機能異常
D	皮膚
	紅斑，皮膚剥脱，乾燥肌，掻痒
G	泌尿生殖器系，性ホルモン
	出血，消化器疾患，血管性浮腫，先天異常，塞栓症
H	全身性のホルモン調節剤，性ホルモンとインスリンを除く
	下痢，鼻炎，腹部膨満，先天異常，胸痛
J	全身性の抗感染症薬
	タンパク尿，下痢，発疹，血尿，肝炎，無顆粒球症
L	抗悪性腫瘍薬，免疫調節剤
	白血球数減少，リンパ球減少症，血小板減少症
M	筋骨格系
	関節痛，消化器疾患，肝機能異常，無顆粒球症，筋膜炎
N	神経系
	神経系疾患，精神障害，運動障害
P	駆虫性薬剤，殺虫剤，忌避剤
	発疹，頭痛
R	呼吸器系
	頻脈，不整脈，低血圧，咽頭炎
S	感覚器系
	苦味，角膜びらん，角膜混濁，筋膜炎，流涙増加
V	その他
	味覚障害，無顆粒球症，健忘，過敏性，パニック反応

(2) 副作用の発症率

利用するデータ「ターゲットタンパク質との相互作用スコア上位第 1 位」「ATC コード第 1 レベル」「副作用発症率」の特徴を分析した．図 2 では，ATC コード毎のスコアと発症率より，発症率があまり高くはないデータであることが分かる．

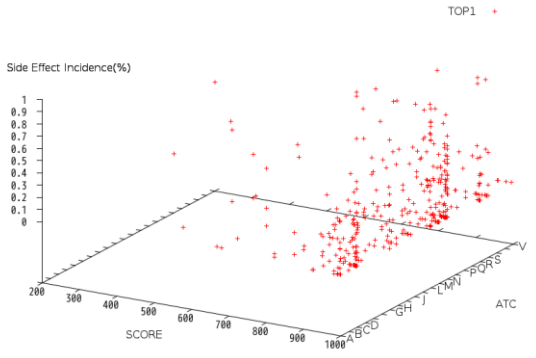

図 2　発症率データの分析

Vol.2014-DBS-160 No.27
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.27
2014/11/18

「タンパク質と化合物（薬）の相互作用スコアの高い化合物（薬）ほど副作用が発症しやすい」という指摘がある [11]. そこで，統計ソフト R [18]を用いて，「ターゲットタンパク質との結合度第1位から第10位」「ATC コード第1レベル（ここでは，アルファベット順に昇順の数字を割り当てた）」「副作用発症率」の相関関係について分析を行った．その結果である変数間の相関を図 3に示す．スコアは，1000点満点による評価である．「スコア」と「発症率」には，正の相関がある傾向を示している．しかし，スコアが900点以上で，副作用発症率が0.1%である薬は，約20%を占めている．また，HIV 融合阻害薬エンフビルチド（enfuvirtide）では，タンパク質との結合度が211点から465点であるが，副作用の発症率は96%である．

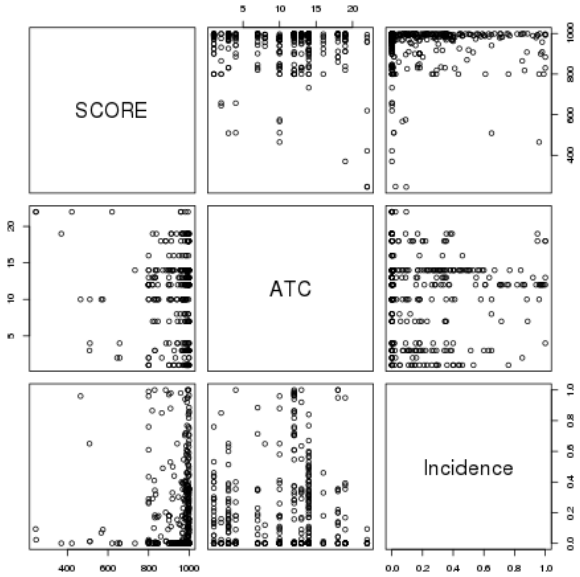

図 3　スコア・ATC 第 1 レベル・発症率の相関

3.4 副作用発症率の予測モデル

(1) PLS 判別分析による 2 値化

図 3より，スコアと ATC コード第1レベルには，やや相関がある．重回帰分析をする際，説明変数の中に互いに相関が高い変数が含まれる場合，通常の最小２乗法では回帰係数の推定精度悪くなることがあり，多重共線性の問題がある．このような問題がある場合，計量化学の分野で開発された PLS（Partial least squares:部分的最小二乗法）回帰が有益な手法である [19][20]. 次に，PLS 回帰式モデルを定義する．

$$\text{予測値 } y' = a_1 * SCORE + a_2 * ACT + b$$

a_1 は説明変数 SCORE の係数，a_2 は説明変数 ATC （ATC コード第 1 レベル）の係数，b は切片（定数）を示す．PLS 判別分析では目的変数を説明する変数（説明変数）を用いて，判別基準を作り，その基準を元に所属グループを判別する．判別基準に PLS 回帰式を用いた線形判別分式を用いて，値が正であるか負であるかで所属するグループを判別する．次に PLS 判別分式を定義する．

$$f(x) = y - y'$$

y は観測値を示す．さらに，PLS 判別分式を用いた判別基準を定義する．

$$\text{If } f(x) \geq 0 \text{ Then } \mathrm{sgn}[f(x)] = 1$$

$$\text{If } f(x) < 0 \text{ Then } \mathrm{sgn}[f(x)] = -1$$

統計ソフト R を用いて，説明変数「ターゲットタンパク質との相互作用スコア第1位」「ATC コード第1レベル」，目的変数「副作用発症率」とした PLS 回帰分析を行い，各係数を求めた．その結果を表 3に示す．

表 3　PLS 回帰式における係数

係数	値
切片	0.27
説明変数 SCORE	0.0001434418
説明変数 ATC	0.000006310492

図 4では，PLS 判別分式の結果を示す．判別分式の値が正であるか負であるかという2値により，所属するグループを「0.1%-40%」「41%-100%」の2段階に分けることが可能である．

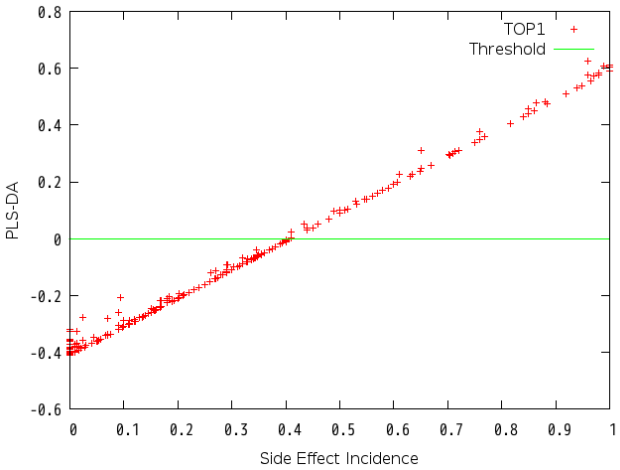

図 4　判別分式による 2 値化

上記の判別分式における正答率は次の通りである（表 4）．表 4より，「ターゲットタンパク質との相互作用スコア第1位」のデータ群を100%の正確さで，２値化することができることが分かる．

Vol.2014-DBS-160 No.27
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.27
2014/11/18

表 4　判別分式の正答率

分類	判別結果 41%-100%	判別結果 0.1%-40%	正答率 (%)
41%-100%	67	0	100
0.1%-40%	0	316	100
合計	67	316	100

(2) **SVM による予測**

PLS 判別分析により得た 2 値データを用いたサポートベクターマシン（SVM）により，化合物とタンパク質の新しい組み合わせにおける副作用発症率を予測する．SVM は，1995 年頃に AT&T の V. Vapnik が発表したパターン識別用の教師あり機械学習方法であり，局所解収束の問題が無い長所がある [21]．現在，2 クラスのパターン識別器としては最も優秀な性能を持つとされている．SVM は，台湾国立大学の Lin らによって作られた SVM のライブラリ LibSVM Version 3.18 を用いた [22]．予測モデルとなる訓練データは，正例データは「41%-100%」データ群，負例データは「0.1%-40%」データ群とする化合物とタンパク質の相互作用スコア第 1 位群である．また，評価データは，化合物とタンパク質の相互作用スコア第 1 位から第 10 位までのデータとした．次にモデルを定義する．

入力空間　$X = \{(SCORE_1, ATC_1), \cdots, (SCORE_n, ATC_n)\}$
出力定義域　$Y = \{1, -1\}$
訓練データ　$S = ((x_1, y_1), \cdots, (x_n, y_n)) = (X*Y)^n$

n はサンプルの数，x_n はいちサンプル，y_n はサンプルのラベルを示す．次に，クラス分類を定義する．

$$f(x) = <w \cdot x> + b$$

w は重み付けベクタ，b はバイアスを示す．さらに，SVM による分類基準を定義する．

If $f(x) \geq 0$ Then $sgn[f(x)] = 1$　（正例）
If $f(x) < 0$ Then $sgn[f(x)] = -1$（負例）

表 5 では，訓練データおよび評価データともに，不均衡データであることを示す．

表 5　データ数

データ区分	訓練データ	評価データ
正例	70	567
負例	342	2760

LIBSVM では，クラス別に重み付けができる．この重み付け係数を利用して，不均衡データを調整する．重み付け係数は次のように算出した．

正例クラス重み付け係数 = 評価データの負例データ数
/訓練データの正例データ数
= 40

負例クラス重み付け係数 = 評価データの正例データ数
/評価データの負例データ数
= 0.2

本モデルのでは，初期 SVM 値設定されている C-SVC (multi-class classification)を用いて，重み付け係数を上記より，次のように設定した．

svm-train **-c 1 -w1 40 -w-1 0.2** training.data

4. 評価

構築した副作用発症率予測モデルの性能を評価するために，訓練データおよび評価データにおける正答率を求める．表 6 および表 7 より，負例クラス重み付け係数 0.2 の時，訓練データおよび評価データともに平均値では一番良い正答率となる．しかし，副作用発症率「41%-100%」および「0.1%-40%」ともに正答率が高く，正例および負例データともに同値がある場合は正例と見なしたいので，負例クラス重み付け係数は，0.19 とする．

表 6　訓練データの正答率

負例クラス重み付け係数	41%-100% (%)	0.1%-40% (%)	平均 (%)
0.19	100	65	83
0.2	100	67	84
0.21	100	67	84

表 7　評価データの正答率

負例クラス重み付け係数	41%-100% (%)	0.1%-40% (%)	平均 (%)
0.19	69	37	53
0.2	22	85	53.5
0.21	37	75	56

5. まとめ

本稿では，PLS 判別分式を用いた 2 値化により，SVM は副作用発症率を分類し予測できるモデルの構築について述べた．PLS 判別分式による分類では，「41%-100%」および「0.1%-40%」ともに，正答率 100%となる．SVM を用いた副作用発症率予測では,「41%-100%」の訓練データは 100%, 評価データでは 69%の正答率を得ることができた．

今後は，病因子遺伝子 [23]や SNP [24]など遺伝子情報と

Vol.2014-DBS-160 No.27
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.27
2014/11/18

疾患の関係から薬の副作用や効能を導き出し，さらに詳細な副作用予測モデルを構築するために，遺伝子情報によりタンパク質や化合物の相互作用に制約を与え，副作用の少ない薬を発見できるような創薬データマイニングを研究開発していく予定である．

謝辞 本研究は，独立行政法人科学技術振興機構 CREST「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化」研究領域として行われた．

参考文献

1) 辰巳邦彦: ドラッグ・リポジショニングと希少疾患イノベーション，政策研ニュース, No. 35, pp.1-9，日本製薬工業協会, March (2012)
http://www.jpma.or.jp/opir/news/news-35.pdf
2) Repurposing Drug
http://www.ncats.nih.gov/research/reengineering/rescue-repurpose/rescue-repurpose.html
3) Xing-Ming Zhao, el.al.: Prediction of Drug Combinations by Integrating Molecular and Pharmacological Data, PLOS COMPUTATIONAL BIOLOGY, Volume 7, Issue 12, December (2011)
http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002323
4) 沼田 稔: DR 研究（既存薬再開発）への期待と課題，医薬ジャーナル， Vol.46, No.5, pp.39-41，株式会社 医薬ジャーナル社, May (2010)
http://www.iyaku-j.com/iyakuj/system/dc8/index.php?trgid=21231
5) 【アメリカ】 増加する薬の副作用報告 2011 年 3 月 28 日，
http://www.ava-net.net/world-news/149-1.htm
6) 病気と遺伝子の関連を解析し，医薬品を生み出すゲノム創薬，
http://www.lifescience-mext.jp/genomme.html
7) 文部科学省オーダーメイド医療実現化プロジェクト，
http://www.biobankjp.org/
8) W. Frawley and G. Piatetsky-Shapiro and C. Matheus: Knowledge Discovery in Databases, AI Magazine, Fall, pp.213-228 (1992).
9) D. Hand, H. Mannila, P. Smyth: Principles of Data Mining, MIT Press, Cambridge, MA, ISBN 0-262-08290-X (2001).
10) Michael Kuhn, et al.: Sytematic identification of protein that elicit drug side effects, Molecular Systems Biology 9, Article number 244 663 (2013)
http://www.ncbi.nlm.nih.gov/pubmed/23632385
11) Yamanishi, Y., Pauwels, E., and Kotera, M.: Drug side-effect prediction based on the integration of chemical and biological spaces, Journal of Chemical Information and Modeling, 52, No.12, pp.3284-3292 (2012)
http://pubs.acs.org/doi/abs/10.1021/ci2005548
12) 津谷喜一郎，五十嵐中，森川 馨: ATC/DDD とは何か-医薬品の合理的使用を目指すものさし-，薬剤疫学 Jpn J Pharmacoepidemiol, 9(2) Dec, pp.53-58 (2004)
http://atc.umin.jp/ATC-DDD.pdf
13) 解剖治療化学分類法
http://ja.wikipedia.org/wiki/解剖治療化学分類法
14) 医療用医薬品の ATC 分類
http://www.kegg.jp/kegg-bin/get_htext?jp08303.keg
15) SIDER 2 Side Effect Resource
http://sideeffects.embl.de/
16) STITCH 4.0.
http://stitch.embl.de/cgi/show_input_page.pl?UserId=OHIBlKK7UnQa&sessionId=aPJvDHSBPj9B
17) Michael Kuhn, et.al.: STITCH 4: integration of protein-chemical interactions with user data, Nucleic Acids Research, November 28 (2013)
http://nar.oxfordjournals.org/content/early/2013/11/28/nar.gkt1207
18) The R Project for Statistical Computing
http://www.r-project.org/
19) 多変量解析
https://upo-net.ouj.ac.jp/tokei/contents/sub_contents/c01_07_00.xml
20) 橋本淳樹，田中豊: PLS 回帰におけるモデル選択多変量解析
http://www.seto.nanzan-u.ac.jp/msie/nas/academia/vol_010pdf/10-03949.pdf
21) Nello Cristianini, John Shawe‐Taylor: サポートベクターマシン入門,共立出版,ISBN 978-4-320-12134-8 (2005).
22) LIBSVM – A Library for Support Vector Machines
http://www.csie.ntu.edu.tw/~cjlin/libsvm/
23) KEGG DISEASE
http://www.genome.jp/kegg/disease/disease_ja.html
24) JSNP DATABASE
http://snp.ims.u-tokyo.ac.jp/