The University of Osaka
Institutional Knowledge Archive

Title	Development of Statistical Software and Implementation of its Models for Operating Pharmaceutical and Genomics Data
Author(s)	周, 怡
Citation	大阪大学, 2020, 博士論文
Version Type	
URL	https://hdl.handle.net/11094/77619
rights	
Note	やむを得ない事由があると学位審査研究科が承認したため、全文に代えてその内容の要約を公開しています。全文のご利用をご希望の場合は、大阪大学の博士論文についてをご参照ください。

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

様式３

論　文　内　容　の　要　旨

氏　名　（　ZHOU YI　）	
論文題名	Development of Statistical Software and Implementation of its Models for Operating Pharmaceutical and Genomics Data （薬学およびゲノムデータ解析用統計ソフトウェアの開発とモデルの実装）

論文内容の要旨

Statistics has great contribution to pharmaceutical and medical research. However, not so many pharmaceutical or medical researchers are familiar with the complicated statistical methodologies and programming. The powerful statistical calculation tools usually require programming skills, such as R. R is regarded as the standard statistical programming languages; however, it lacks a graphical user interface (GUI) that appeals to various users. Current GUIs built on top of R, such as EZR and R-Commander, aim to facilitate R coding and visualization, but most of the functionalities are still accessed through a command-line interface (CLI). To assist researchers of medicine and pharmacy in running the most routines in fundamental statistical analysis, an interactive GUI, MEPHAS, was developed to support various web-based systems that are accessible from laptops, workstations, or tablets, under Windows, macOS (and IOS), or Linux. In addition to basic statistical analysis, advanced statistics such as the extended Cox proportional hazard (CoxPH) model and dimensional analyses including partial least square regression (PLS-R) and sparse partial least square regression (SPLS-R) are also implemented in MEPHAS. An executable R package mephas was also implemented with the user-friendly GUI available under various computing environments.

Users can perform data analysis tasks such as managing data, analyzing data, and visualizing the results step by step without intensive statistical software training. MEPHAS covers adequate pharmaceutical statistics including statistical probability distributions and hypothesis testing with various data types. It also provides advanced statistical methods such as analysis with regression models and dimensional analyses. In addition, MEPHAS extends regression models to cover random or interactive effects and enable the prediction of the new dataset. MEPHAS made up for the lack of the AFT model in most statistical software and provided two types of time-to-event data analysis. It is also the first web-based application to produce dynamic results and 3D plots for PLS-R and SPLS-R methods.

PLS has been widely used in chemoinformatic data analysis. In the past decades, PLS also gained lots of applications in bioinformatic data. With the development of high-throughput technologies, gene expression data have become easier to obtain. Such data are enhancing our understanding of cancers and some other

intractable diseases. Great concerns have been raised on using gene expression data to predict cancer patients' survival time. However, as the most popular model to predict the survival times, the CoxPH model does not work on data with a great number of variables. To deal with this issue, researchers have developed various partial least squares (PLS) algorithm in CoxPH model to reduce the dimension of variables for survival prediction. As the extension of PLS and CoxPH model, the model, FiPLSCox, which combines forward interval PLS (FiPLS) and the CoxPH, was proposed and can be used to predict cancer patients' survival outcomes.

PLS applied with CoxPH models for genomic data censored time were firstly discussed on both simulated gene expression data and real-world cancer data. The proposed FiPLSCox had competitive prediction performance compared with the previous PLSCox model and the prediction performance was assessed by the time-dependent AUCs and time-dependent prediction errors. The results indicated that this proposed FiPLSCox model had good performance for classifying new patients into clinically relevant high-risk or low-risk groups based on the gene expression and survival data from previous patients. Additionally, if the baseline hazard could be known, the final outcome from FiPLSCox can be used to predict patients' survival probabilities, and furthermore to assist clinical physicians in making clinical decisions in the early diagnosis.

様式 7

論文審査の結果の要旨及び担当者

氏　名　（　Zhou Yi　）			
論文審査担当者		（職）	氏　名
	主 査	教授	高木　達也
	副 査	教授	上田　幹子
	副 査	教授	大久保　忠恭

論文審査の結果の要旨

統計学は医薬品や医薬学研究に必須の知識であり、ツールとなってきたことは疑う余地がない。しかし、その複雑な統計手法に精通している医薬学研究者は多くない。複雑な統計計算には、通常、Rなどのアプリケーション実行スキルが必要だが、GUI（グラフィカル・ユーザー・インターフェース）に欠けているため、多くの実験薬学研究者に、習熟に至る時間的余裕はないのが現状である。そこで申請者は、医薬学の研究者が基本的な統計解析を実行するのを支援する目的で、Windows、macOS（およびIOS）、またはLinuxの下で、ラップトップ、ワークステーション、またはタブレットからアクセス可能な、WebベースでRを稼働させるシステム、MEPHASを開発した。

基本的な統計解析に加えて、拡張Cox比例ハザード(CoxPH)モデル、部分最小二乗回帰(PLS-R)、スパース部分最小二乗回帰(SPLS-R)を含む多変量解析法など、医薬学で特に必要性が高いと思われる高度な統計解析法もMEPHASに実装した。また、ダウンロードしてオンサイトで実行可能なRパッケージmephasも実装されており、様々な計算環境で利用可能なユーザーフレンドリーなGUIを備えている。

MEPHAS、またはmephasにより、ユーザーは、統計ソフトのトレーニングを受けることなく、データの管理、データの分析、結果の可視化などのデータ分析作業を行うことができる。MEPHASは、様々なデータ型を用いた確率分布や仮説検定など、使用頻度の高い医薬統計を網羅しているだけでなく、回帰モデルを用いた分析などの高度な統計手法も提供している。さらに、MEPHASは拡張回帰モデル（ランダムモデルや交互作用効果を含む）、2種類のイベント間時間データ解析、PLS-R、SPLS-R法の動的結果と3Dプロットを生成する最初のウェブベースのアプリケーションとなる。

PLSは計量化学分野で広く利用されているが、近年、計量薬学の分野でも利用され始めてきた。特にハイスループット技術の開発により、遺伝子発現データの入手が容易になるに伴い、その大量データの解析が必要となってきた。このようなアレイデータの応用範囲の一つに、難治性悪性腫瘍患者の生存期間を予測することが挙げられるが、同時に懸念も指摘されている。生存期間を予測するための最も一般的なモデルとしては、CoxPHがあるが、大量の変数からなるデータでは機能しない。この問題に対処しようと、生存予測のための変数の次元圧縮を考慮、CoxPH－PLSアルゴリズムが開発されてきたが、そのままでは予測性、頑健性に問題があった。申請者は、PLSとCoxPHモデルの拡張として、前方区間PLS(FiPLS)とCoxPHを組み合わせたモデルFiPLSCoxを提案、悪性腫瘍患者の生存予後予測に利用することを可能にした。

遺伝子アレイデータに対するCoxPHモデルを用いたPLSの適用について、まず、シミュレーションで作成された遺伝子発現データと実際の悪性腫瘍データの双方について議論した。提案したFiPLSCoxは従来のPLSCoxモデルと比較して競合的な予測性能を有しており（予測性能は時間依存AUCと時間依存予測誤差によって評価）、FiPLSCoxモデルが、過去の患者の遺伝子発現データと生存データに基づいて、新規患者を高リスク群と低リスク群に分類するのに良好な性能を有していることが示された。

以上、医薬学に有用な統計解析ツールの開発と、新規な生存時間解析モデルの開発を行ったことにより、博士（薬科学）の学位論文に値するものと認める。