

京都大学若手人材海外派遣事業 ジョン万プログラム  
研究者派遣プログラム

成果報告書

提出日：平成 29 年 2 月 22 日

【基本情報】

○申請者

採 択 年 度：平成 28 年度

部 局 名 等：医学研究科社会健康医学系専攻

職 名：特定教授

氏 名：田中 司朗

研究課題名：医療系データベースを用いた因果推論のための生物統計手法の開発と応用

○渡航先

国 名：米国

研究機関名：The University of North Carolina at Chapel Hill

研究室名等：[研究室名] Department of Epidemiology

[職名等・氏名] Professor Alan Brookhart

渡 航 期 間：平成 28 年 10 月 14 日～28 年 11 月 13 日、平成 28 年 11 月 21 日～29 年 1 月 25 日  
(97 日)

○渡航期間中の出張

出 張 先：第 22 回日本薬剤疫学会学術総会（みやこめっせ）

目 的：日本薬剤疫学会で成果の一部を報告

期 間：平成 28 年 11 月 14 日～28 年 12 月 24 日

京都大学若手人材海外派遣事業 ジョン万プログラム  
研究者派遣プログラム

【成果】

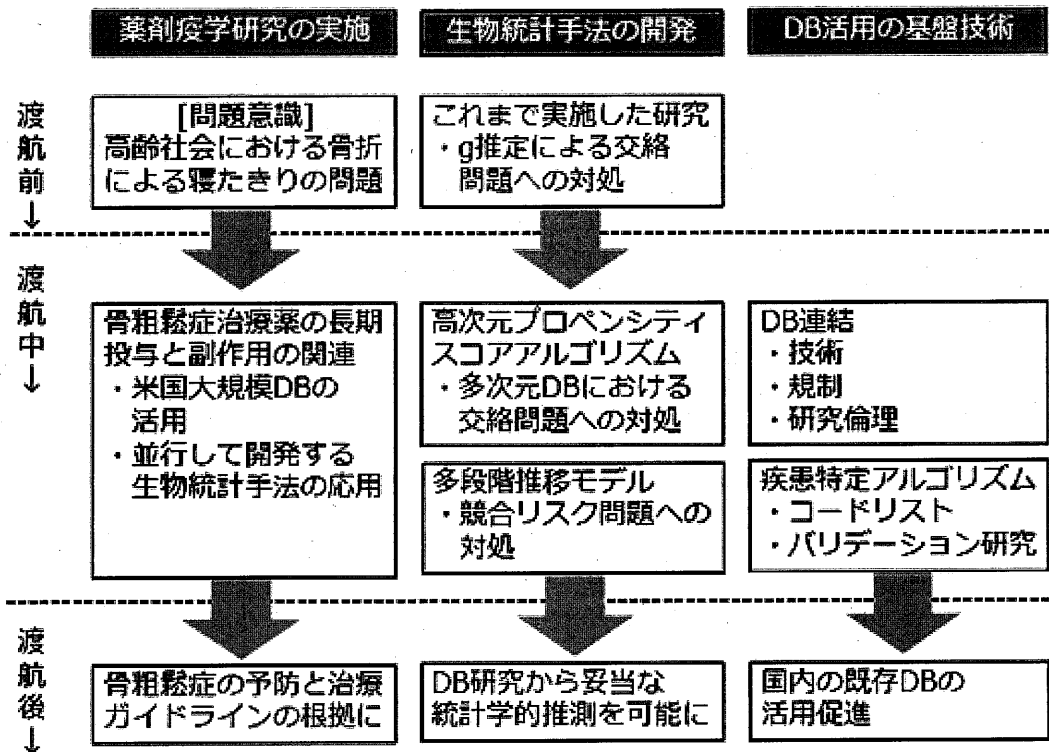
○プロジェクトの成果及び今後の展開

・研究概要

本研究計画は、図1の概要を示されているように、生物統計手法の開発、薬剤疫学研究の実施、DB活用の基盤技術の習得の三本柱からなる。

まず、受入研究者であるノースカロライナ大学 Alan Brookhart 教授・同大学の Jason Fine 教授との共同研究として、医療系データベースを用いた薬剤疫学研究のための生物統計手法の開発を行う。すなわち、交絡調整・競合リスクという薬剤疫学研究で生じる方法論的課題を数学的に定式化し、それを解決するための統計手法を提案した後に、数学的な証明・導出、コンピューターシミュレーション、実データへの適用を通じて新しい統計手法を開発し、その性能を評価する。実データへの適用は、骨粗鬆症治療薬研究など具体的な薬剤疫学研究を実施し、現実のデータセットを用いて新規統計手法の有用性を評価するものである。更に、米国で用いられているデータベース連結に関する技術・規制・研究倫理やデータベースから疾患を特定するために用いられているアルゴリズムについて習得する。海外渡航を通じて開発・習得した技術は、論文という学術成果にとどまらず、我が国の既存データベースの利活用のための基盤となる。

図1. 研究計画の概要



京都大学若手人材海外派遣事業 ジョン万プログラム  
研究者派遣プログラム

・国際共同研究の立上げ・ネットワークの構築

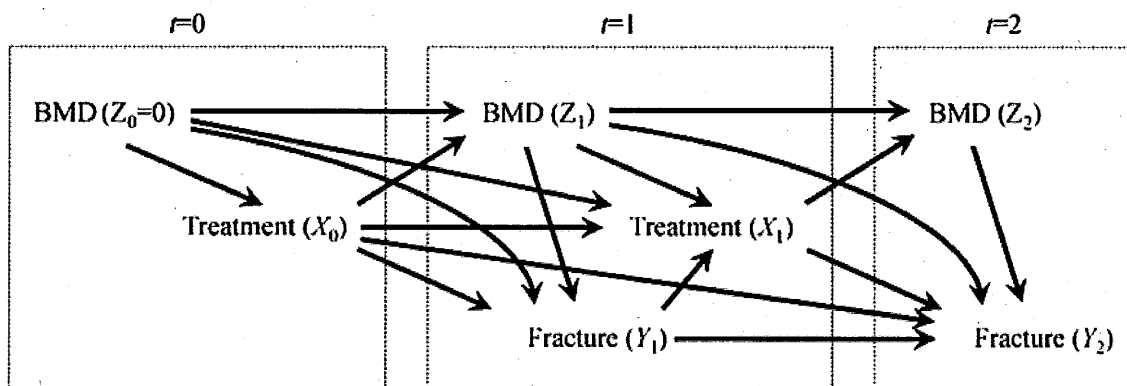
生物統計学の研究の一般的な流れは、研究で生じる現実の問題を数学的に定式化し、それを解決するための統計手法を提案した後に、数学的な証明・導出、コンピューターシミュレーション、実データへの適用を通じて新しい統計手法を開発し、その性能を評価するものである。

本研究テーマの一つは、「交絡」と呼ばれる因果関係のゆがみを補正するための統計手法である。図2に、原因（骨粗鬆症治療）、結果（骨折）と交絡因子（骨密度）の関係を示す。骨粗鬆症治療薬は骨密度が低い重症な患者を選んで処方される。従って、治療を受けている患者や骨折高リスクであり、単純に治療あり群となし群を比較すると、見かけ上治療あり群で骨折リスクが高いという解析結果になってしまう。申請者は先行研究で、交絡問題に対処するためにg推定という統計手法を開発した（Epidemiology 2007、図1）。g推定とは、治療と傾向スコアの残差をとり、それが潜在結果変数と独立になることを前提に、推定方程式を構成する数理的手法である。しかしながら、骨折以外に死亡など複数の重篤な疾患が生じる状況下の統計解析では、別のイベントが発生したときには関心のあるイベントが観察できない競合リスクという問題が生じる。競合リスクの存在下でのg推定の推定方程式を構成するよい手法は提案されていない。

この問題を数学的な問題として解決するための統計手法を提案するため、ノースカロライナ大学 Alan Brookhart 教授・同大学の Jason Fine 教授とのゼミ形式での議論を行った。具体的に議論された内容は、因果推論・競合リスク・推定方法に関する先行研究のレビュー、pseudo-observations 法という新規推定方法の統計学的性質に関する数学的な検討、コンピューターシミュレーションの設定とその結果の解釈、実データへの適用とその解釈であった。

また、ノースカロライナ大学は研究基盤として医療系データベースが極めて充実しており、その取り扱いに関する技術や研究倫理・規制が先進的である。例えば我が国の現状では困難なことの一つが複数のデータベースの連結である。米国では、診療報酬請求、診療記録、処方記録、がん登録情報など様々なデータベースを個人レベルで連結させ、分析に用いることが行われている。その際にはしばしばプライバシー保護の問題が生じる。そこで、米国医薬品食品局が行っている薬剤の安全性に関するプロジェクトでは、コモンデータモデルと呼ばれる分散型の新しいデータベース統合モデルが用いられている。これらのデータベース連結に関する技術や研究倫理・規制について習得するため、大学院生・研究者向けのゼミである Pharmacoepi Seminar（毎週月曜）、Causal Inference Labo（毎週月曜）、PS Meeting（毎週月曜）に出席した。更に、研究集会である International Dose Effect Alliance Workshop 2016（11月2日）、James Robins Seminar（11月3日）に出席し、それぞれ放射線疫学・因果推論に関する最新の知見を得た。

図2. 骨粗鬆症治療、骨折と骨密度（BMD）の継時的な関係



京都大学若手人材海外派遣事業 ジョン万プログラム  
研究者派遣プログラム

・国際共著論文の投稿・発表等の状況、国際学会等での発表状況 [予定を含む]

論文

1. G-estimation of structural nested mean models for competing risks data using pseudo-observations. Biometrics. Shiro Tanaka, Alan Brookhart, Jason Fine. (予定)

学会発表

1. G-estimation of structural nested mean models for competing risks data using pseudo-observations. ISPE's 10th Asian Conference on Pharmacoepidemiology. Shiro Tanaka, Alan Brookhart, Jason Fine. (予定)
2. G-estimation of structural nested mean models for survival data using pseudo-observations. Kyoto Biostatistics Seminar. Shiro Tanaka. (予定)
3. データベース研究・患者報告アウトカム. 第22回日本薬剤疫学会学術総会. 田中司朗. (2016年11月20日)

・在外研究経験によって習得した能力等

Alan Brookhart 教授、Jason Fine 教授とのゼミ形式での議論では、準備として、基礎となる論文や教科書をレビューした後に、pseudo-observations 法という新規推定方法やその統計学的性質の数学的な導出を主に行った。これを行うためには数学的なトレーニングが必要であり、それを通じて、pseudo-observations 法やその周辺領域である生存時間解析の数理に習熟した。また、継続的に議論を行った経験は、英語でのコミュニケーション・プレゼンテーション・ディスカッションスキルの向上に結び付いた。本プロジェクト終了後は、平成31年10月に行われる ISPE's 11th Asian Conference on Pharmacoepidemiology の Local Host 事務局長を務める予定であるが、英語スキルの向上や海外の薬剤疫学者との緊密な関係は、今後の国際的な研究活動に寄与するところが大きいと考えている。

・在外研究経験を活かした今後の展開

本プロジェクト終了後の平成29年2月1日から、AMED 生物統計家育成支援事業「京都大学大学院における臨床統計家育成のための教育カリキュラム標準化のための研究開発」にて特定教授として雇用されることとなった。臨床統計家とは統計学を基礎とする臨床試験方法論のプロフェッショナルのことであり、この事業は臨床統計家の育成を目的とする教育プログラムである。このような背景から、本プロジェクトによるノースカロライナ大学チャペルヒル校での滞在経験を通じて、公衆衛生大学院における生物統計教育について洞察を行った。我が国と欧米の生物統計教育の違いは三つあると考えられる。第一に、コースワークの充実（科目数やコースで課せられる課題の量）が挙げられる。臨床試験で用いられる生物統計手法は極めて専門分化しており、質・量ともに充実したコースワークを行わなければ実務を行うための学問的素養を身につけることはできない。第二の違いとして、メンタリングの質が高いことがある。欧米の指導教員の多くは生物統計学の研究者として一流であるため、例えば学位論文のテーマ選定一つとっても日本よりレベルが高いと感じられる。そして、第三の違いとして、学部教育と公衆衛生大学院で行われる生物統計学専門教育との連携がある。我が国の生物統計学教育には、理工学部などの数理系学部と公衆衛生大学院が設置されている大学院医学研究科では分野が異なるため、学部教育と生物統計学専門教育の接続が悪いという構造的な問題がある。これらの知見は今後の教育活動に活用し、欧米の生物統計学教育と同水準のメンタリング・コースワークの提供を通じて、海外に通用するだけの能力を持った生物統計家を育成したいと考えている。

京都大学若手人材海外派遣事業 ジョン万プログラム  
研究者派遣プログラム

英文成果報告書

○申請者情報

部 局 名 : Department of Pharmacoepidemiology, Graduate School of Medicine and Public Health, Kyoto University, Kyoto, Japan

職 名 : Associate Professor

氏 名 : Shiro Tanaka

研究課題名 : Development and application of biostatistical methods for causal inference using medical databases

渡 航 期 間 : October 14, 2016 to November 13, 2016, November 21, 2016 to January 25, 2017

○渡航先情報

国 名 : The USA

研究機関名 : The University of North Carolina at Chapel Hill

研究室名等 : Department of Epidemiology

受入研究者名 : Professor Alan Brookhart

○渡航報告

Competing risks frequently arise in medical research. A common approach to the competing risks problem is to examine the effects on the cumulative incidence functions of stroke which is relevant even in the presence of competing risks. Another issue is that the relations between exposures and health outcomes are often subject to bias due to confounding. In particular, statistical methods for adjustment for time-dependent confounders have not been established in the presence of competing risks. In this project, I conducted a methodological work from a statistical viewpoint on this common problem in medical research collaborating with Professor Alan Brookhart and Professor Jason Fine at the University of North Carolina at Chapel Hill.

The basic idea of our research is to derive a g-estimator of structural mean models using pseudo-observations. Suppose a cohort study of an i.i.d. sample of  $n$  subjects which aims to estimate the effects of exposure  $A$  on the time to failure  $T$  with  $J$  mutually exclusive failure types and right-censoring time  $C$ . Let  $\{X_i, \Delta_i, \Delta_i \varepsilon_i, A_i, L_i, i=1, \dots, n\}$  be the observed data, where  $X_i = \min(T_i, C_i)$ ,

京都大学若手人材海外派遣事業 ジョン万プログラム  
研究者派遣プログラム

$\Delta_i = I(T_i \leq C_i)$ ,  $\varepsilon = 1, \dots, J$  is the cause of failure,  $L$  is the covariates and  $I(\cdot)$  is the indicator function. We assume that there exist potential outcomes  $\{T_i^a, X_i^a, \Delta_i^a, \Delta_i^a \varepsilon_i^a\}$  that represent the outcomes of the  $i$ th subject that would be seen were the  $i$ th subject to have an exposure level denoted by  $A_i = a$ . Consider linear structural mean models of the potential outcomes:  $\psi_0 + L\psi_1 = E(Y^a | A, L) - E(Y^0 | A, L)$  where  $\psi^T = \{\psi_0, \psi_1^T\}$  is the causal effect of exposure  $A$  and the interactions with  $L$ .

Suppose that it is possible to compute the nonparametric estimator of the cumulative incidence function of interest (the first failure type)  $\hat{C}_1(t)$  at the end of the study. Let  $\hat{\theta}$  and  $\hat{\theta}_{-i}$  be the estimated cumulative incidence function from the entire dataset and its leave-one-out version obtained by eliminating the  $i$ th individual. The pseudo-observation for the  $i$ th subject is defined as  $\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$ . It can be shown that this quantity has a useful property, that is,  $E(\hat{\theta}_i) = E(Y_i | A_i, L_i)$ .

Define  $H_i(\psi) = \hat{\theta}_i - (\psi_0 + L_i\psi_1)A_i$  for estimation of the linear structural mean models.

It follows from the property of pseudo-observations that

$$E\{H_i(\psi) | A_i, L_i\} = E(Y_i | A_i, L_i) - (\psi_0 + L_i\psi_1)A_i$$

From the consistency assumption and the linear structural mean models, we have that

$$\begin{aligned} E\{H_i(\psi^*) | A_i, L_i\} &= E(Y_i | A_i, L_i) - (\psi_0^* + L_i\psi_1^*)A_i \\ &= E(Y_i^0 | A_i, L_i) \end{aligned}$$

where  $\psi^*$  is the true value of  $\psi$ . If the no unmeasured confounders assumption holds,  $A_i$  and  $Y_i^0$  (therefore  $H_i(\psi^*)$ ) are mean independent conditional on  $L_i$ , and a class of moment estimating functions for  $\psi$  is derived:

$$U(\psi) = [d(A, L) - E\{d(A, L) | L\}]w(L)[H(\psi) - q(L)]$$

京都大学若手人材海外派遣事業 ジョン万プログラム  
研究者派遣プログラム

where  $d(A, L)$  and  $w(L)$  are any vector-valued functions with the dimension of  $\psi$ , and  $q(L)$  is an arbitrary scalar function. For any choices of  $d(A, L)$ ,  $w(L)$  and  $q(L)$ , a consistent estimator of  $\psi$  is obtained by solving the resulting estimating equations.

In this project, we formulated structural mean models of causal effects on the cumulative incidence functions of potential outcomes and developed the novel estimators of structural mean models based on pseudo-observations. We also extended the framework to proportional hazards structural mean models which represents causal subdistribution hazard ratios over the entire time span and structural nested mean models for time-varying exposure. We also provided the finite-sample performance of the proposed estimators in simulated datasets and applied the proposed methods to a real dataset from a cohort study.