厚生労働省科学研究費補助金(循環器疾患･糖尿病等生活習慣病対策総合研究事業)
「追跡終了後コホート研究を用いた共通化データベース基盤整備とその活用に関する研究」
研究報告書

疫学追跡終了後コホートデータの共通利用（アーカイブ化）の際の死因データ利用に関する検討

研究分担者　大橋　靖雄　　中央大学理工学部人間総合理工学科生物統計学
研究協力者　原田亜紀子　　東京大学大学院医学系研究科生物統計学

研究要旨

死因情報を付与した形でのデータアーカイブ化が難しい現状を鑑みて、現制度下での運用案を提案した。一つ目はアーカイブセンターなどに死因情報以外のデータを集約管理し、必要に応じて従来通りの死因照合作業を実施し、死因を付加したデータセットを作成し解析を行う方法である。二つ目は、データを保持者のもとに置いたまま、必要と判断した情報だけを選択的に共有させる分散型ネットワークによる方法である。提案1の運用案を複数のコホート研究から構成される国内循環器疫学研究（JALS）に適用したところ、死因照合率は 99.7%であり、照合作業の技術的側面、作業手順化の面で問題はなく、運用案の一つになりうると考えられた。提案2については、医療（臨床）、疫学研究で先行している分散型ネットワークの事例を収集し、本検討事例である疫学追跡終了後コホートデータの共通利用の場面に応用可能な事例があるか検討を行った。疫学共同研究における分散型ネットワークを用いたデータ連携の例などから、リモート環境で様々な統計解析を柔軟に行えること、ローカルサイト（大学、研究所、医療機関を想定）へアクセスを行う上でのセキュリティ、さらには、人口動態統計の二次利用をローカルサイト単位で各々申請し、死因を付与したデータを統合して用いることや、施設外部の PC 上の RAM への一時的な書き出し（保存はされない、電子データの「一時的蓄積」）の可否など、法令解釈の問題にも留意する必要があると考えられた。疫学研究追跡終了後のコホートデータの共通利用に関して、提案した二つの方法で運用していくことは可能と考えられるが、実際に運用する場合には、前者においてはデータ利用の規約、データ利用の物理的環境、サポート体制、後者は、人口動態二次利用申請による死因データ利用の法令解釈などの課題があると考えられた。

A. 目的

米国では、National Heart, Lung, and Blood Institute （NHLBI）などの公的研究資金で実施した臨床試験や疫学観察研究のデータは、Biologic Specimen and Data Repositories Information Coordinating Center （BioLINCC） に対して追跡終了後一定の期間、条件を充足した後に研究データを登録し公開するよう進められている（https://biolincc.nhlbi.nih.gov）。BioLINCC に収集されたデータは、使用申請を行うことにより、研究利用でき、研究データの二次利用により新たな研究や若手研究者の教育・研究機会が創出されている。翻って我が国の現状をみると、厚生労働科学研究費補助金を受けた研究では、National Bioscience Database Center (NBDC) 等へのデータ提供が求められ、バイオサイエンス基礎研究でのデータアーカイブ化には進展はみられるが、個人の健康情報を含む疫学観察（コホート）研究領域では、未だデータの二次利用環境は整っていない。このように疫学観察研究において、データの共有化、オープンデータ化が進まない背景には、コホート研究が様々な研究費で維持されており、米国NHLBIなどに代表される大型の研究費を受けた先へのデータ提供という単純な構図になりえないこと、多くのコホート研究では主要なアウトカムとなる死因情報を人口動態統計調査の二次

利用申請によって得ており、死因を付与した状態でのデータ公開が難しい点、研究企画時点から、データアーカイブ化を見据えた倫理的諸問題の対応、調査票やデータベース構造の標準化の準備されていないなどが原因として考えられる。そこで本研究では、死因情報を付与した形でのデータアーカイブ化が難しい本邦の現状を鑑みて、現制度下で疫学追跡終了後コホートデータの共通利用を行うための運用案を提案する。

B. 方法

1. 死因情報を外したアーカイブ環境を想定し、必要時に死因を人口動態二次利用申請し、アーカイブセンター (データセンター)にて照合・集計・解析を行う運用

国内大規模循環器疫学研究（JALS）を例に、研究を統括する中央研究事務局が人口動態二次利用申請し、提案する運用例にならって死因を照合することで生じる問題点を考察し、提案例をアーカイブセンターで運用する際に想定される課題を検討した。

1）死因データを必要時に連結させる方法の提案

JALS では、ローカルコホートから追跡調査データ（生存・死亡（死因はなし）、発症）の提供があり、死因の同定については中央事務局(データセンター)で人口動態の二次利用申請を行い一括で照合作業を行っている。この JALS で採用しているシステムを例にして、研究コンソーシアムに属する研究者が、このデータを利用した研究を計画し、人口動態の二次利用申請で死因情報を得ることによって、センターのスタッフ（必ずしも専門知識を必要としない）により簡便に照合作業が行えるような環境を検討した。

2）人口動態統計の二次利用申請

Japan Arteriosclerosis Longitudinal Study (JALS)は、国内 58 市町村、8 職域の計 11 万人を追跡している。この JALS コホートのうち、職域コホートを除き、研究開始（2002 年　ただし一部コホートは 1999 年）から 2012 年 12 月までの異動状況が、住民基本台帳（住民票）情報により確認されている追跡対象者で期間中に死亡した 7,137 件を死因照合の対象とした。総務省による「統計法第 33 条の運用に関するガイドライン」を参照した上で、厚生労働省大臣官房統計情報部企画課審査解析室に人口動態調査二次利用の申請を行った。申請作業と並行し研究事務局において 2012 年 12 月までの死亡例について、性別、生年月日、死亡日、死亡時の居住地（市町村コード）のリストを作成した。

3）提案例に従った照合作業

人口動態調査の使用許可がおりた後、性別、生年月日、死亡年月日、死亡時の居住地（市町村コード）を照合変数とし、人口動態調査データを用い原死因を確定した。

2. 分散型ネットワーク国内外の医療（臨床）データ連携、疫学共同研究利用などの先行事例を収集し、疫学追跡終了後コホートデータの共通利用に活用できるかどうか検討を行う

C. 結果

1. 死因情報を外したアーカイブデータを利用する運用案

1）データ提供からアーカイブ化まで

提案するデータ利用基盤の概略を[図 1]に示した。研究コンソーシアムに参加する各研究が、基本データ（生活習慣、検査データなど）と死因を除いた追跡データをアーカイブセンターに提供する**（①データ提供）**。アーカイブセンターでは、基本データベースと追跡データベースを分けて構築しておき、基本データベースは原則登録時から修正なしの状態、追跡データベースは、今後の追跡継続に応じて更新できる構造とする**（②アーカイブ化）**。追跡データは、今後の死因照合作業で必要となる「死亡地（市町村）」、「死亡日」、「生年月日」、「性別」を含むよう設計した。

2）死因情報の申請

コンソーシアム内の研究者（あるいは一定の条件を設け、研究グループ外の研究者も可能とするか）

が、このデータベースを使用する研究を計画し、死因情報を得るために厚生労働省に対して人口動態調査二次利用申請を行うとした。**(③死因情報申請)**。

3）照合作業から研究データセット作成

承認後に提供を受けた死因情報をアーカイブセンター内で、「死亡地（市町村)」、「死亡日」、「生年月日」、「性別」をキー変数として、保有する追跡情報と照合し**（④データ照合)**、死因を付与した一時的な解析データセットを作成し、研究計画に基づいた解析に使用する**(⑤解析用データセットの作成)**。研究終了後は死因情報を削除(抹消)し、厚生労働省に利用後報告を行うという流れである**(⑥利用報告、死因情報の削除)**。死因を付与したデータをアーカイブせずに、研究を計画するたびごとに、この①から⑥の作業をアーカイブセンターで手順化し対応するという運用方法の提案である。

4）JALS を例にした上記作業の検証

2015 年 1 月 29 日付で利用許可がおり、JALS 対象地域の市町村で 1999 年 1 月 1 日から 2012 年 12 月 31 日までに発生した死亡の調査票情報を受領した。

① 提供を受けた人口動態データの読み込み

提供を受けるデータの構造(提供変数)に応じて変更は必要となるが、データの構造と読み込み用の SAS プログラムを作成した。

②研究全体の市町村(コード)をリストアップした一覧表を読み込み（SAS プログラム）

③追跡調査データベースの対象データから死因照合を行いたい死亡者を抽出する作業（SAS プログラム）

④照合作業（SAS プログラム）

上記①と③で作成されたデータセットについて、「死亡地（市町村コード)」、「死亡日」、「生年月日」、「性別」でマッチマージしていく。JALS の対象者で、職域コホートと死亡調査データが確定していないコホートを除き、死亡が特定できていたのは 7,137 件であった。性別、生年月日、死亡年月日、死亡時の居住市町村名を照合変数とし、人口動態調査データと一致がみられたのは 7,099 件（99.5%）であった。

⑤照合例、未照合例のリスト化（SAS プログラム→Excel ファイル）

死因一致例の中に複数の候補例(一意に決まらない例)が存在する対象がいないか、また未照合例の情報確認を行うために該当者のリストアップを行う必要がある。これらを SAS から Excel ファイルに出力するプログラムを作成した。複数の候補例が見つかったのは 5 件あり、いずれも東日本大震災の被災地域での死亡者で、死因が同じであったためその死因で照合を行った。死因が照合できなかった例は 33 件であり、このうち 14 件については、以前に JALS が行った死因照合作業において既に未照合が判明

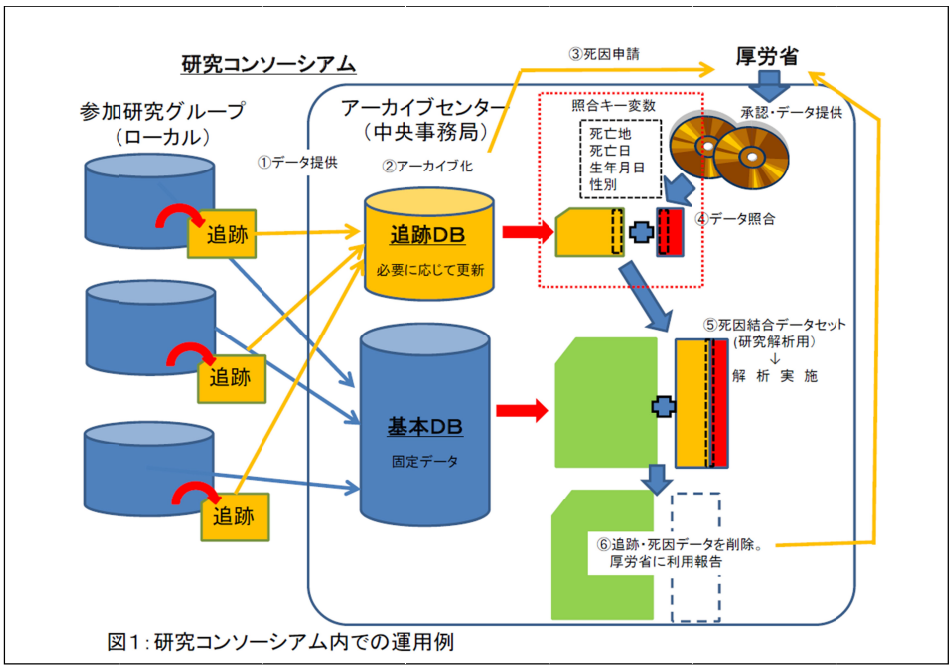

図1：研究コンソーシアム内での運用例

しており、各コホートに対して死亡時情報を確認したがいずれも情報に誤りのなかった例で、人口動態統計作成の過程で入力間違い等が発生した事例と判断した。このため、今回照合出来なかった例は、実質として 19 件（0.27%）であった。

２．分散型ネットワーク国内外の医療（臨床）データ連携、疫学共同研究利用などの先行事例収集と、疫学追跡終了後コホートデータの共通利用についての検討

１）国内外の分散型ネットワーク先行事例（医療・臨床データ連携（交換）の収集

①国内：Standardized Structured Medical record Information eXchange （SS-MIX）

厚生労働省電子的診療情報交換推進事業(Standardized Structured Medical record Information eXchange)で策定された『電子的診療情報を他システムとの交換や地域医療連携で利用するために、診療情報を標準的な形式で蓄積・管理するデータとして保存できる領域』の仕様のことである。SS-MIX では、これらの医療情報を「標準化ストレージ」というツールに医療情報を標準化した形式で格納・蓄積することにより、複数ベンダー間・複数システム間の相互運用性を高めることを目的としている。標準化ストレージの構造は、コンピュータの一般的なファイル格納形式と同様に階層化されたフォルダのディレクトリー構造を用いている。電子カルテシステムなどではほとんどの業務で個々の患者を軸として診療情報が格納されているため、標準化ストレージにおいても患者にひも付く各種の情報をフォルダの階層構造にルールを決めて格納している。それ以外の項目について研究間での統合をはかるために、拡張ストレージを設け統合することで、標準化ストレージを核にし、施設間連携を構築し、診療情報の研究利用、地域医療連携、災害時連携などを目指している)。

②米国：Sentinel Initiative データ交換モデル

米国では 2008 年より、Food and Drug Administration Amendments Act (FDA 改革法)によって FDA 承認医薬品および医療製品の安全性モニタリングのため、既存のデータベースを用いたアクティブ・サーベイランスである Sentinel Initiative が実施されることとなった。Sentinel Initiative ではデータインフラと手法の開発を目的として, Harvard Pilgrim Health Care Institute が中心となって医療データベースを所有する数十の医療機関が協力して Mini-Sentinel も実施され, 各データストレージは分散したままで, コモンデータモデルに従って共通プログラムで解析するシステムが構築されている。

② 米国：Health Information Exchange (HIE)

CDX（Crossflo Data Exchange®）ソフトウェアは、Crossflo 社が開発したデータ交換ソフトウェアで、HL7、GJXDM (Global Justice XML Data Model)、NIEM (National Information Exchange Model)、EDXL、CAP、NCPDP など主要医療データシステムおよび国家的なデータ標準、業界標準に準拠したデータ共有方式を実装している。

本システムの医学領域の応用事例としては、モンタナ州および連邦政府の要請を受けて、国立医療情報科学センター の協力のもと、モンタナ州の 4 つの病院の ED（救急科）と州保健福祉局 (Montana Department of Public Health and Human Services) とをデータ接続するモンタナ州医療情報交換システムプロジェクト (Montana Health Information Exchange Pilot Project) があげられる。CDX によるシステムで、ほぼリアルタイムでの医療情報の交換を実現し、異種の医療データソースを迅速および効率的に接続する疾病サーベイランスシステムが構築されている。

２）国際共同疫学研究等での分散型ネットワークによるデータ連携例

ViPAR:Virtual Pooling and Analysis of Research Data (Karter KM et al. *Int.J.Epidemiol.* 2015)

ViPAR は、International Collaboration for Autism Registry において、国際共同研究を実施す

る6つのサイトデータを統合的に解析することを目的に開発された（Karter KM et al. *Int.J.Epidemiol.* 2015）。多くの医学研究では、倫理的、法的な理由から保有するデータの管理と保持を各サイトにおいて行う必要があるが、本研究では、解析の目的で一時的に「Virtual pooling」サイトに移動することは許可されたことから、一か所にデータを持続的に収集することなしに、物理的に離れたリモートサイトのデータを仮想的(一時的)に集約し分析を可能とするシステムが開発された。ViPAR のリモートサイトとマスターサーバの関連の関連を図 4 に示したが、ViPAR のマスターサーバは、リモートサイトの PC へ接続されており、研究データはそれぞれリモートサイトで管理されている。利用者は Web 上の analytical portal からアクセスし、データをリモート PC からマスターサーバの RAM 上に抽出し、仮想的にデータをプールして解析を行い、終了後は保存することなく消去する。このため外部にデータを収集することなく多施設データを統合した解析が可能となっている。

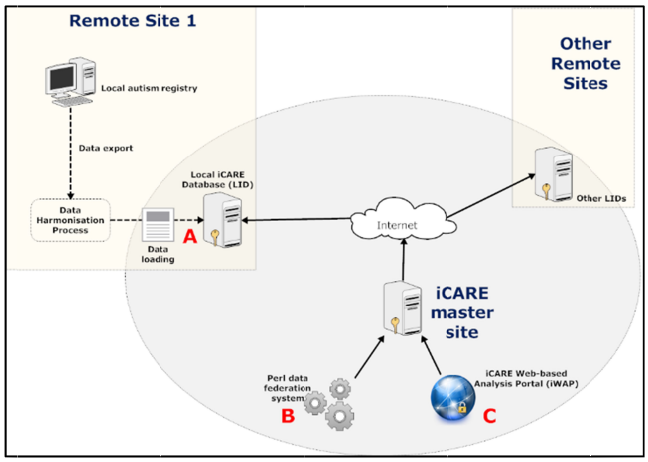

図 2 ViPAR のリモートサイトとマスターサーバの関連（Karter KM et al. *Int.J.Epidemiol.* 2015）

ViPAR の環境のうち、リモートサイトのデータベースは、データ storage として MySQL サーバを、マスターサイトとの通信のため SSH サーバが導入され構成されている（いずれもオープンソース）。実際に解析を行うマスターサーバは、ViPAR daemon と web-based のポータルで構成されている。ViPAR daemon は、ログ出力、統計パッケージのアクセスとコントロール、ローカルサーバから抽出したデータの統合作業を担う。web-based のポータルは、解析、データマネージメントを行うインターフェイスであり、新規の解析スタート画面、解析アウトプット画面、コードのマネージメント画面の三つからなる。ViPAR を用いた具体的な解析の流れは、以下の通りである。

① 解析用インターフェイスを開き変数やサイトを選択
② 統計パッケージの種類（R, SAS, Stata）を選択し、テキスト入力部分に解析のための syntax を入力しサブミットする
③ リモートデータベースにデータのリクエストが送られ、マスターサーバの RAM に読み込まれ、バーチャルプールされる（これらのデータは保存されることはない）。バーチャルプールデータセットは、選択された統計パッケージに読み込まれ、解析が行われる
④ ファイルマネージャー画面で、解析の進行状況、完了状況が確認でき、すべての結果とログがダウンロード可能になっている。

ViPAR のプログラム、マニュアル等は下記で公開されている
http://bioinformatics.childhealthresearch.org.au/software/vipar/

D. 考察

1．死因情報を外したアーカイブ環境を想定し、必要時に死因を人口動態二次利用申請し、アーカイブセンター（データセンター）にて照合・集計・解析を行う運用

本検討では、死因情報を外したアーカイブ環境を想定し、必要時に死因を人口動態二次利用申請し、アーカイブセンターにて照合・集計・解析を行う運用例を提案し、国内大規模循環器疫学研究（JALS）で、実際に運用し死因照合を行った。JALS は各コホート研究（ローカルサイト）データを中央事務局で統合して実施している研究であることから、この研究で行う死因照合作業の実際が、本検討で想定している研究データアーカイブの環境づくりに対して参考になる点が多いように思われた。死因データの提供を受けてから実施する照合作業については、JALS の研究進捗にあわせて、統計プログラム（SAS）

を用い、データの読み込み、Excel などへの帳票出力に至るまで一連の作業の手順化をすすめてきた。多くの作業をルチーン化してきており、大部分の課題は解決してきているが、本提案例をアーカイブセンターで運用する際に予想される課題について検討する。

1）厚生労働省から受領する統計フォーム（データ形式）に大幅な変更がないことを前提とする

本検討で提案した仕組みは、人口動態二次利用申請により提供される死因データの構造に大幅な変更がないことを前提としている。構造の変更を伴わない小規模の変数追加等の変更であれば、受領後の確認作業を手順化すること（変数表の確認作業やプログラム修正箇所の明示）で、アーカイブセンタースタッフ、疫学研究に従事する研究者等で十分対応できる作業と考えられる。

2）死因照合率はコホートから提出される照合キー変数の精度に依存している

死因照合の一致率については、各コホート研究から提出される照合キー変数（死亡地、性別、死亡年月日、生年月日）の情報精度に依存している。JALS では、自治体の住民基本台帳などの確認後の情報に基づき照合作業を実施しているが、未照合となる例は、提供情報に何らかの誤りがあり、情報の再確認後に照合可能となる例が大部分である。一方で、提供情報に誤りがないにもかかわらず照合例を検索できない例が少なからず存在しており、これらは性別、生年月日、死亡日について、人口動態統計作成の過程で発生した何らかの入力ミス等が原因として考えられる。しかし、本検討で想定しているのは追跡終了後コホート研究のデータアーカイブ化であるので、これまでの研究実施の過程（各研究での死因照合作業などの過程）でこれらの変数は十分確認されてきており、情報の精度については、進行中の研究に比べて十分高く、こうした問題の発生は少ないと考えられる。

3）死因付与データが使用できる環境が限定される

提案する方法では、死因付与後のデータセットの利用場所がアーカイブセンター（あるいは申請書に記載した研究者の所属する機関）に限られることになる。したがって、研究の解析も同様にこの利用場所に限られることになり、データ解析を行える環境（物理的な環境、統計家の配置等）についても検討する必要がある。また、死因データの申請者の所属機関で実施する場合には、アーカイブデータの外部利用の規約等の整備も必要があるといえる。

2．分散型ネットワーク環境を利用し、疫学追跡終了後コホートデータの共通利用が可能かどうか

分散型ネットワークが開発され、利用されている背景には、倫理的規定などからデータ使用の場が制限され物理的に施設外に出せないが、何らかの方法でデータ利用を促進したいというニーズが存在する。疫学研究でのデータ連携では、医療（臨床）データ連携で必要とされるような即時的なデータ連携、多種多様なデータをそのまま連携したいというニーズは高くはない。一方で、以下のような医療データ連携にはない、疫学研究で重視すべき項目が考えられる。

① 様々な統計解析を柔軟に行えること

医療（臨床）データ連携のように、定型化された集計作業（例：Mini sentinel）が中心ではないので、ViPAR のような各種統計パッケージ等と連動して運用可能であるかどうかは重視すべき点であると考えられる。ただし、ViPAR については、RAM 上で解析を行い、外部にデータ保存が行われないというだけであり、施設外にデータが出るということの定義や解釈が利用の際の要件となってくる。

② ローカルサイト（大学、研究所、医療機関を想定）へアクセスを行う上でのセキュリティ

昨今、施設によっては外部からのアクセスに対して制限を設ける場合も多くなっている点を考慮すべきである。

③ 人口動態統計の二次利用に関連した法令解釈

・ローカルサイト単位で各々申請し、死因を付与したデータを統合して用いることの可否

・ViPAR での運用のように、施設外部の PC 上の RAM への一時的な書き出し（保存はされない、電子データの「一時的蓄積」）の可否

などが問題として考えられる。

文部科学省文化審議会著作権分科会等でも機器

利用時・通信過程における一時的なデータ蓄積については議論されており、「一時的固定（複製）」については、次のように整理されている（コンピュータに該当する部分抜粋）。

ア．瞬間的・過渡的な蓄積であり「複製」ではないもの
・処理装置(CPU)の読込み
・ビデオ RAM への書込み

イ．一時的固定（複製）のうち、「複製」と判断すべきものではないもの
・主記憶(RAM)への蓄積
・補助記憶のドライブキャッシュ注釈
・CPU における１次キャッシュ２次キャッシュ

ウ．一時的固定（複製）のうち「複製」と判断すべきもの
・主記憶(RAM)への蓄積（常時蓄積）

このような解釈は、技術動向を見極めて判断されるものでもあり難しい課題であるといえる。

一施設にデータを集約しない分散型ネットワークによるデータ連携（交換）は、本検討の既存研究データの連携に限ったことではなく、今後新たに疫学研究を行う際にも検討していく必要がある。このような方法を活用していくことで、共同研究の促進や研究のマネージメントの効率化・生産性の向上が期待できるほか、信頼性の高いデータ収集へつながっていくものと考えられる。

E. 結論

死因情報を付与した形でのデータアーカイブ化が難しい本邦の現状を鑑みて、現制度下での運用案を提案した。提案１の運用案を複数コホート研究から構成される国内の循環器疫学研究に適用したところ、死因照合率は 99.7%であり、照合作業の技術的側面、作業手順化の面では問題はなく、運用案の一つになりうると考えられた。想定するアーカイブセンター等で実際に運用を行う際には、データ利用の規約、データ利用の物理的環境、サポート体制などについては検討すべき必要があると考えられた。提案２については、疫学共同研究における分散型ネットワークを用いたデータ連携の先行事例などから、様々な統計解析を柔軟に行えること、ローカルサイト（大学、研究所、医療機関を想定）へアクセスを行う上でのセキュリティ、さらには、人口動態統計の二次利用をローカルサイト単位で各々申請し、死因を付与したデータを統合して用いることや、施設外部の PC 上の RAM への一時的な書き出し（保存はされない、電子データの「一時的蓄積」）の可否など、法令解釈の問題にも留意する必要があると考えられた。

F.研究発表

1.　論文発表
2.　学会発表
　いずれもなし

G. 知的財産権の出願・登録状況

（予定を含む。）

1. 特許取得
2. 実用新案登録
3.その他

いずれもなし