[Case Report]

# データマネジメント効率化を目的としたプログラミング言語 R の研修プログラムの構築

# Establishment of a Programming Training of R Beginner-Oriented Specialized for Effective Data Handling

■

伊藤　典子　　鳥居　　薫　　西岡絵美子　　齋藤　明子　　堀部　敬三

**ABSTRACT**

***Background/Objectives***：Nagoya Medical Center datacenter has been supporting more than 80 clinical trials since 2003. Moreover, it has increased the number of supports continuously. Therefore, to establish a system which enables the datacenter to support more clinical trials efficiently is its significant theme. Especially, periodical monitoring report making consumes about 20% resources out of all datacenter works. This study sets to examine if application of an IT system leads to efficiency after revealing the issues during making reports, as its goal.

***Methods***：At first, we grasped the current conditions and revealed the issues. Next, we considered IT application, established and performed a training program for IT application. We also examined the effectiveness of the training program by using test data.

***Results***：It was found out that data handling with Excel took much work time during periodical monitoring report making. We selected R to shorten the time and established a programming training of R beginner-oriented specialized for data handling for data managers.

To examine the effectiveness, we measured both current work time and work time adopting R by using test data and made comparison between them. Accordingly, we performed a simulation. As a result, we obtained an implication that great resource reduction can be realized by adopting R.

（**Jpn Pharmacol Ther 2016；44 suppl 2：s155-160**）

**KEY WORDS**　R, data handling, data manager, clinical trial, programming training

## はじめに

名古屋医療センターでは臨床研究品質確保体制整備病院として，各種疾患領域の臨床試験に対し高品質かつ効率的な支援体制を構築してきた。名古屋医療センター臨床研究センターのデータセンターは2003 年に臨床研究支援を開始し，現在約 80 試験以上の臨床研究支援の実績を有しており，症例登録，症例報告書（CRF）の設計，データマネジメント，中央モニタリングや解析結果レビューなどの幅広い業務を担当している。臨床研究支援数が増加しており，限られたリソースで品質を維持したまま効率的

独立行政法人国立病院機構名古屋医療センター臨床研究センター
Noriko Ito, Kaoru Torii, Emiko Nishioka, Akiko Saito, Keizo Horibe: Clinical Research Center, National Hospital Organization Nagoya Medical Center

により多くの臨床試験を支援できる体制を整備することが重要な課題である。特に，逸脱や有害事象情報の抽出を目的として定期的に作成している定期モニタリングレポートは，データクリーニングや，データの不整合などについてクエリー発行も実施しており，データセンター業務の約2割のリソースを消費している。定期モニタリングレポート作成における問題点を明らかにし，ITシステムを導入することで，効率化につながるか検討することを本研究の目的とした。

## 対象と方法

当データセンターにおけるデータマネジメント業務を対象とし，以下の4つのステップに分けて検討した。

### 1　現状把握・問題点の明確化

多大なリソースを消費している定期モニタリングレポート業務に焦点を当て，業務内容やプロセスをデータマネージャーより詳細にヒアリングし，問題点を明確化する。

### 2　IT導入の検討

費用対効果も含め，1で確認された問題の解決につながる可能性のある適切なIT技術を選択する。

### 3　研修プログラムの構築と実施

ITを導入し現場で利用できるようにするための教育研修プログラムを構築し，データマネージャーに教育する。

### 4　研修プログラム導入の有効性検討

#### 1）テストデータを用いた評価

教育研修プログラム導入による有効性を評価するため，受講者に対しテストデータを用いて以下の2通りの作業方法でデータ加工を行い，作業時間を計測し比較した。

計測時間1：従来実施しているExcelファイルでの作業時間

計測時間2：研修で学んだR言語プログラムでの作業時間

評価に用いるテストデータは，研修内容を考慮して作成する。また，3名の受講生は独立してテストを実施する。

#### 2）シミュレーションによる評価

シミュレーションに基づき，教育研修プログラム導入による有効性を評価した。

## 結　　果

### 1　現状把握・問題点の明確化

#### 1）現状把握

定期モニタリングレポートに掲載する逸脱や安全性情報の一覧は，名古屋医療センターで独自開発された電子的データ収集システム（electronic data capture: EDC）のPtosh（patient data organizing system）[1]からダウンロードされたcsvファイルをもとに作成している。通常，電子症例報告書（eCRF）は，ビジットごとに作成しており，Ptoshからのダウンロードcsvファイルは，ビジットごとにeCRFでの取得項目が含まれるかたちで作成されるため，複数ファイルから成る。そこでまず，複数のcsvファイルを1つのExcelファイルの各シートにまとめる。そのExcelファイルに逸脱抽出や集計に必要な関数を埋め込み，データ抽出や逸脱条件などを組み立てる。ここで利用される関数は四則計算のような簡単な関数から，シート間のLOOKUP関数などの検索関数，IF関数を入れ子にした複数条件を設定した関数が存在し，これらを組合せた複雑で長い関数として利用することが多かった。逸脱抽出や集計する量も試験で取得するcsvファイルのデータ量に依存するが少なくない。1試験あたりcsvファイルは数ファイルから数十ファイルから成り，1つのcsvファイルの変数も数十から数百項目となっている。データ量が大きい試験は13メガバイトの取得項目が格納されたExcelファイルを扱っている。大量なデータ量と複雑なExcel関数処理を実施してレポートを作成しているため，エラー防止の運用として内容確認を目視のダブルチェックを実施している。

なお，当データセンターでは，定期モニタリングレポート作成を年に1〜2回実施しており，データクリーニング用，レポート作成用など，逸脱や安全性情報の一覧作成といったデータ抽出業務は，各定

期モニタリングレポート作成1回のなかで数回ずつ繰り返し作業していた。

**2）問題点**

逸脱や安全性情報の一覧を作成するためのExcelファイルを作成するにあたり，Excel関数を1セルごとに組み込み，関数をコピーしながら作業するため，手間がかかり，エラー発生率は低くないと考えられた。また，エラー発生予防を目的としたダブルチェックには，当該Excelファイル作成とほぼ同時間がかかっていることがわかった。データ量が多い試験では，作業中のExcelファイルのフリーズやファイル故障も頻繁にあり，再作成のため，作業時間をさらに増大させているものもあった。これらより，定期モニタリングレポート業務に多大なリソースを消費する原因はExcelを利用したデータハンドリングの煩雑さによる作業時間の膨大化であることがわかった。

## 2　IT導入の検討

Excelでのデータハンドリングにかかる膨大な作業時間を短縮するため，この作業工程を代替できるIT導入を検討し，プログラミング言語としてR言語を選択し利用することにした。R言語は，ニュージーランドのオークランド大学のRoss Ihaka氏とRobert Gentleman氏により開発が開始され[2]，世界中で利用されている統計解析用のフリー（無料）ソフトウェアである。R言語を選択した理由はデータ量によらない効率的なデータハンドリングが可能であること，オープンソースのソフトウェアであり多くのボランティアにより協力にサポートされメンテナンスと拡張が行われていること，コスト面ではフリーであるためソフトウェアにかかる費用の増大がないことがある。また，長期的に期待できる側面として，統計解析用の言語のため定期モニタリングレポート作成のみの利用でなく，シニアデータマネージャー以上の知識があれば，グラフや集計などの簡単な統計解析も実施することができると考えた。

## 3　研修プログラムの構築と実施

**1）研修プログラムの構築**

①研修により習得可能なスキルの目標値設定

研修によって受講者のスキルをどのレベルまで上げられるか検討した。R言語によるプログラムを利用すれば定期モニタリングレポート作成業務でExcelを利用している作業をすべて代替することもできる。しかし，当院のデータマネージャーはExcelを利用するスキルには長けているが，他のIT技術には明るくないメンバーも多くプログラミングは未経験であり，変数の概念を含めたデータベースの基本も含め教育する必要があった。そこで受講者のIT知識量も含めて検討した結果，データハンドリングの基本となるファイル入出力，マージ，ソートと簡単な条件式をR言語でプログラムできるようになることを目標とした。

②研修の内容

現状把握と研修により習得可能なスキルの目標値設定から，以下の研修内容・方針を決定した。

- ファイル入出力（csvファイルの概念），マージ，変数の追加・削除・並替え，変数名の変更，ソート，条件式を含める。基本的なデータハンドリングだけでも現状のExcelの手作業と比較しても格段の効率化が行えることが想定された。
- R言語のみの講義ではなくデータベースの概念やcsvファイルの構造など必要なIT知識を補足する。
- インストールと環境設定も研修内容に含める。プログラムコードは問題ないが，環境設定により，うまく作動しない場合があることを知識として含める。
- 特にデータフレームのデータハンドリングに重点を置く。利用しているExcel関数を代替するのはデータフレームでまかなえる。
- プログラム演習を多く取り入れる。プログラムは知識だけ頭に詰め込んでもスキル習得にはならない。実際に手を動かし，エラーで動かない状況や考えたとおりの結果にならないことを試行錯誤した経験が重要となると考えた。
- 可読性は重視と考え，コメントと利用法は含める。
- 開発環境はRコンソール，Rエディタを利用。
- 研修時間以外の宿題は基本的にはなしとする。

一方，教養としてプログラムを教えることが目的ではなく，また，プロのプログラマー養成するのが目的でもないことから，内容に含めないことを整理

するのも大切なステップであると考え，研修内容に含めない点を，以下のとおり決定した。

・R 言語はオブジェクト指向のプログラミング言語であるが，プログラマーを養成するのが目的ではないため，オブジェクト指向の概念は研修に含めない。
・効率的なプログラムを作成するには最適化されたアルゴリズム，メモリなどを考慮した処理時間などがありベクトルの処理を考慮したプログラムが理想であるが，今回の目的はデータハンドリングであり，対象者はプログラミング初心者のため高度なスキルは含めないこととした。処理時間がかかるシミュレーションならば，メモリ消費やベクトルを考慮したプログラムアルゴリズムは必要かもしれないが，ベクトルや行列を含めると R 言語の勉強以前に数学の基礎知識から再度復習させる必要があること，また必要なデータハンドリングはデータフレームを習得できれば目的を満たせるため学習範囲に含めない。
・プログラム自体の汎用性や効率性をすぐに求めていない。

以上の点は，データセンター全体が R 言語をツールとして利用し効率化・標準化に成功し，組織や運用が整備された場合，次のステップでは検討課題に入る可能性がある。

研修内容に含める点・含めない点を前述のように考慮し，プログラミング経験のない受講者のため，教育研修プログラムを構築した。構築した研修プログラムの概要は**表 1** に示す。データハンドリングにかかわる作業を R 言語の利用で効率化し，業務のツールとして利用できるスキルをつける内容とし，全 12 回より構成した。

③講義用資料の作成

講義用の資料は PowerPoint にて自作した。一般的に R 言語のユーザーは統計解析を行うユーザーがメインである。そのため，セミナーや市販本は統計解析を行うための R 言語の本であり，データハンドリングを主としたものがない。本研究の目的，受講者，業務内容に一致するものを自作することにした。

④研修の受講者の選定と方法

研修名は「R トレーニング」とし，定期モニタリングレポートの作成経験があるデータマネージャー 3 名をパイロット的に受講生一期生として実施した。データマネジメント経験が 10 年以上のシニアマネージャー 1 名と約 2 年のデータマネージャー 2 名であった。講義では R 言語を演習しながら進めるため，受講者はノートパソコンを持参し受講した。構築した研修プログラムは段階的に知識を習得するため，講義で疑問が残ったままでは次の講義の理解はさらに苦しい。そのため，受動的な講義では習得が難しいと考え，参加者が主体であることを重視し，疑義があれば逐次発言してもらい些細な疑問点も解決しながら進める対話形式とした。そのため，構築された研修プログラムは各回 1 時間を想定していたが，実施するのに平均すると約 1.2 時間かかった。特にマージに関する講義が 2 時間と長くかかった。

**表 1　R 言語 研修プログラム**

| 回数 | メインタイトル | 内容 |
|---|---|---|
| 第 1 回 | インストール | R の概要，インストール，環境設定 |
| 第 2 回 | 基本計算 | R console，四則演算，変数，combine，平均，最大，最小 |
| 第 3 回 | エディタ操作 | R エディタ，コメント |
| 第 4 回 | データフレーム | データフレーム，データエディタ，欠損値 |
| 第 5 回 | ファイル読込み 1 | 作業ディレクトリ，ファイル読込み（テキストファイルと csv ファイルの概念も含める） |
| 第 6 回 | ファイル読込み 2 | |
| 第 7 回 | データハンドリング | データ抽出，簡単な計算，ソート（概念も含める） |
| 第 8 回 | マージ | 結合，マージ（概念も含める） |
| 第 9 回 | ファイル出力 | ファイル出力 |
| 第 10 回 | プログラム練習 1 | ファイル読込→データハンドリング→ファイル出力を含めた総合演習 |
| 第 11 回 | プログラム練習 2 | |
| 第 12 回 | プログラム練習 3 | |

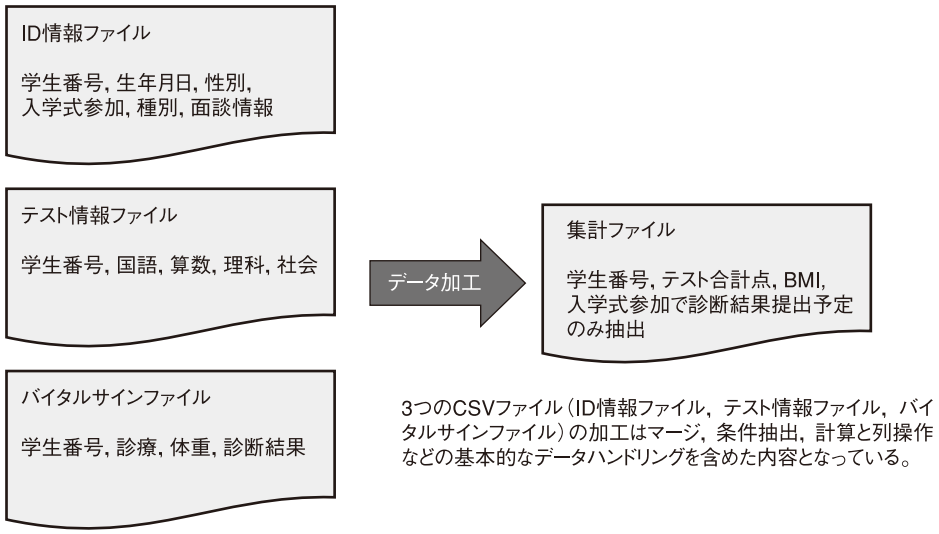


**図 1　テストデータのデータ構造**

### 4　研修プログラム導入の有効性検討

#### 1）テストデータを用いた評価

①テストデータの作成とテスト作業内容の設定

研修プログラムの有効性評価を行う際に用いるテストデータは，研修の内容である csv ファイル入出力，マージ，変数の追加・削除・並替え，変数名の変更，ソート，条件式と欠測値を利用する構造とし，3 つの csv ファイルを加工して 1 つの csv ファイルを作成する作業を想定して作成した（**図 1**）。データ構造は維持したまま，データ内容を変えて 2 種類のデータセットを作成し，テストを 2 回実施させた。

②テストの結果

テストの結果を**表 2** に示す。

従来行っていた Excel を利用した作業（計測時間 1）では，1 回目の平均作業時間は 18 分であった。また 2 回目以降テストデータを更新して同様に作業しても同じだけの時間がかかった。

研修で学んだ R 言語プログラムでの作業（計測時間 2）では，1 回目の平均作業時間はプログラム作成を含めて平均 30 分であった。また，更新されたテストデータを用いた 2 回目の作業にかかる時間は平均 1 分であった。

従来の Excel を利用した作業に比し，R 言語プログラムでの作業は，1 回目が Excel 作業時間の 1.7 倍，2 回目は 0.1 倍の時間という結果であった。

#### 2）シミュレーション

テストの結果を利用し実際にどのくらいのリソースが効率化されるか予想した。たとえば，以下のような仮の臨床研究を 1 つ想定する。

**表 2　テストデータを用いた作業よる研修前後の作業時間の比較**

| | | 受講生 1 | 受講生 2 | 受講生 3 | 平均 |
|---|---|---|---|---|---|
| 計測時間 1（研修前）：従来法の Excel を利用した作業（何度実施しても同じだけ時間がかかる） | | 10 分 | 15 分 | 30 分 | 18 分 |
| 計測時間 2（研修後）：R 言語を利用した作業 | 1 回目 プログラム作成含む | 30 分 | 20 分 | 40 分 | 30 分 |
| | 2 回目 | 2 分 | 1 分 | 1 分 | 1 分 |

■仮の臨床研究

・試験期間：5 年間

・定期モニタリングレポートの作成は年 2 回実施する。レポート作成時のデータクリーニング（データに関するクエリー発行と確認），およびレポート作成のため，1 回レポート作成時に同様のデータハンドリング作業が 3 回発生する。

・従来の Excel を利用したデータハンドリング作業時間は 1 回あたり 3 時間かかるとする。

仮の臨床研究の想定で定期モニタリングレポートに要する時間は，現行の Excel での作業では 3 時間×3 回×年 2 回×5 年分＝90 時間である。R 言語

プログラムでの作業時間は，**表 2** のテストデータを用いた作業による作業時間に基づき，1 回目は Excel 作業時間の 1.7 倍（5 時間），2 回目は 0.1 倍（3 時間）かかることになる。したがって，R 言語プログラムでの作業時間は 1 回目の定期モニタリングレポート作成作業は 5 時間＋0.3 時間×2 回＝5.6 時間，2 回目以降は 0.3 時間×年 2 回×4 年分＝2.4 時間となり合計 8 時間となる。R 言語を利用すれば約 0.5 人月のリソース削減が見込めることがわかった。

## 考　　察

データセンターでの定期モニタリングレポート業務を効率化するために，R 言語の教育研修システムを構築し有効性を検討した。テストデータを用いた本システムの有効性評価より，データハンドリングに R 言語を導入した場合，従来の Excel を利用した作業に比べて，初回のプログラム作成に時間がかかるが，2 回目以降の作業は更新されたデータセットに同じプログラムを再実行するだけであり短時間で実施できることがわかった。定期モニタリングレポート作成のたび，データクリーニング用やレポート作成用に同様のデータハンドリングを複数回繰り返していることを考慮すれば，データハンドリングに R 言語を適用することにより，2 回目以降の作業について大幅な効率化が期待できると考えられた。そもそもプログラムの再利用性は定期モニタリングレポートのようなデータハンドリングの繰り返し作業に適しており，IT 導入の成功例であると考えられた。また，今回は，R 言語を習得したばかりの初心者のためプログラム作成時間に時間がかかったが，実際の現場での利用の経験を積み，研修の強化，プログラミング技術の向上やプロセス管理の実施により，さらに大幅な効率化が予期待される。シミュレーション結果より，Excel による作業に比べ，R 言語プログラムでも作業は 1 試験あたり 0.5 人月のリソース削減が得られることがわかり，当データセンターでは約 80 臨床試験を支援していることから，大幅なリソース削減が図れることが予想された。

データセンターの全データマネージャー約 15 名への R 言語研修の実施を目標とし，2016 年 8 月時点で第 2 期生 2 名，第 3 期生 4 名受講済みとなった。第 2 期生以降の教育には第 1 期生を講師に加えており，受講済データマネージャーの知識の充実につながることを目指している。一方で，現場への導入にあたっては，運用範囲や手順書など利用環境の整備なども必要となる。また，R 言語利用者によるスキルの格差や実務で発生する問題点の対策にフォローアップ研修も必要となっている。

また，今回は効率化を重視して検討したが，質の観点からの評価や検討も必要であり，実務での効率化と質の面を考慮した生産性も検討すべき課題である。さらに長期的には，要約統計量や簡単なグラフなどの統計解析もデータセンターでサポートできるようにすることも課題である。

## 結　　論

データセンター業務の効率化のため，IT の導入を検討した。世界中で汎用される統計解析用フリーソフトウェアの R 言語を選択し，現場のニーズに直結する機能に特化した研修プログラムを構築し実施した。テストデータを用いた有効性検討，およびこれに基づくシミュレーション結果より，大幅なリソース削減が示唆される結果となった。

**［謝　辞］**

本稿の執筆にあたり，独立行政法人国立病院機構名古屋医療センター臨床研究センターのデータセンターの皆様に謹んで感謝の意を表します。

## 文　　献

1）齋藤俊樹，齋藤明子，近藤修平，永井かおり，西岡絵美子，堀部敬三．臨床研究中核病院における臨床試験データの電子化への取り組み．レギュラトリーサイエンス学会誌 2015; 5: 61–71.
2）リゲス U（著），石田基広（翻訳）．R の基礎とプログラミング技法．丸善出版; 2012.
3）The R Project for Statistical Computing. https://www.r-project.org/