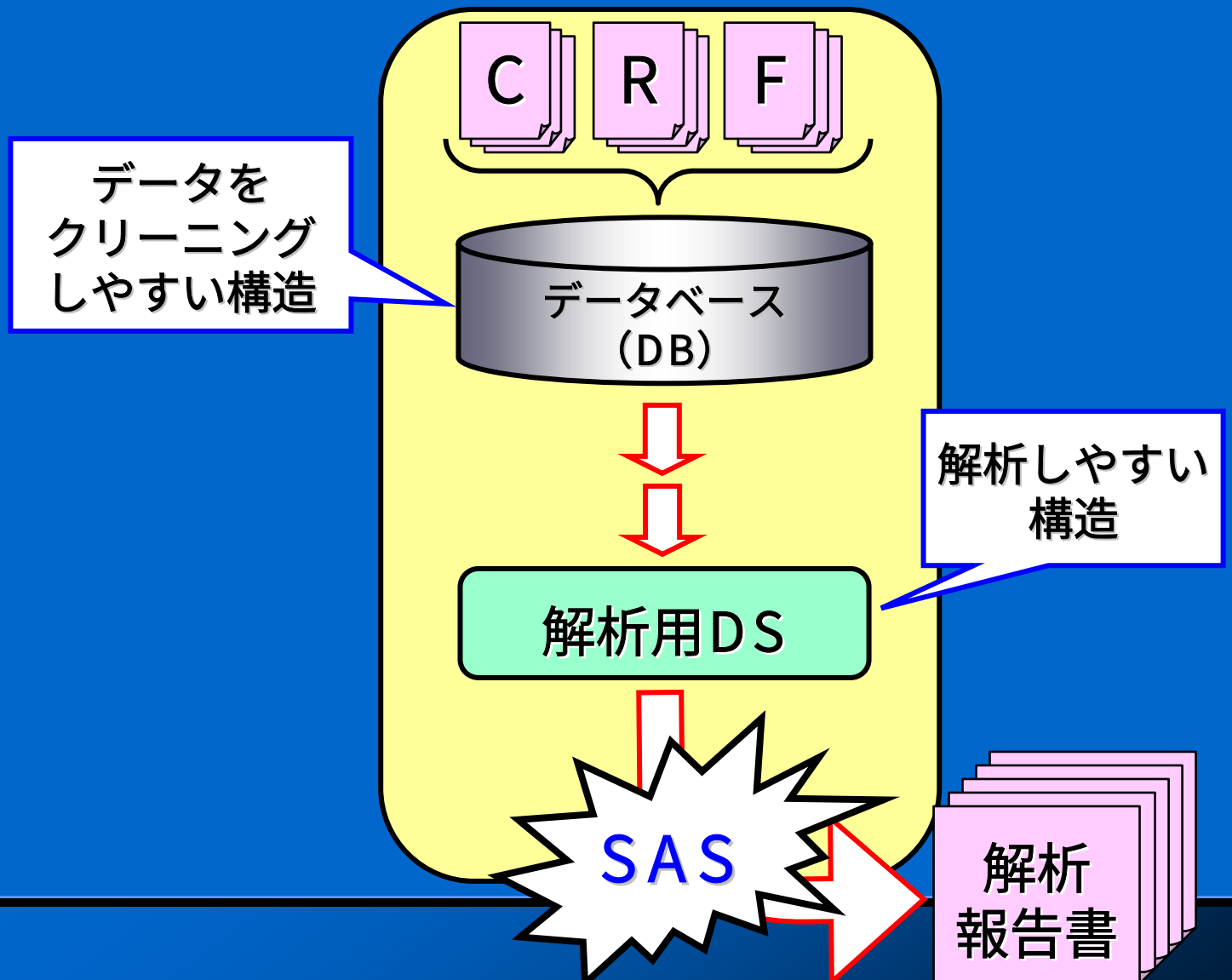
ダブルプログラミングによる解析用データセットの作成

○山本祐史[1]・益田隆史[2]・菅波秀規[2]
データマネージメント課[1]・統計解析課[2]
臨床解析部
興和株式会社

解析用データセットの必要性



ＤＢから解析用ＤＳを作成する過程

わずかなエラーが重大な誤りをもたらす可能性がある

- ＣＲＦ収集段階
  - →副作用用語の付け間違い
  - →個別症例のエラー
- 解析用ＤＳ作成段階
  - →適格例を抽出するプログラム
    適格性条件の等号を忘れる
  - →複数症例のエラー

例） Age <= 65 と Age < 65

データ収集には莫大なコストがかかっている

品質管理が極めて重要

効率的な品質管理を選択したい

品質管理手法

バリデーション
された
プログラムの利用

作成された
解析用DSの
目視による比較

ダブル
プログラミング
によって作成
された解析用DS
のCOMPARE
プロシジャ
による比較

品質管理手法

バリデーション
された
プログラムの利用

作成された
解析用DSの
目視による比較

ダブル
プログラミング
によって作成
された解析用DS
のCOMPARE
プロシジャ
による比較

症例報告書の項目名
DBの変数名やDBの構造
①試験ごとに異なる
②一部標準化されていない

プログラムは試験ごとに
バリデーションしなければならない
→効率良い品質管理？

品質管理手法

バリデーションされたプログラムの利用

作成された解析用DSの目視による比較

ダブルプログラミングによって作成された解析用DSのCOMPAREプロシジャによる比較

目視による比較を２回以上行うことは集中力を保つことが困難

- 修正のたびに目視が必要
- 品質管理記録の作成
  →効率的な品質管理？

品質管理手法

バリデーションされたプログラムの利用

作成された解析用DSの目視による比較

ダブルプログラミングによって作成された解析用DSのCOMPAREプロシジャによる比較

解析用DSの比較は COMPAREプロシジャ	比較のためのコスト　0 品質管理記録作成　0 プログラミングコスト→大

目的

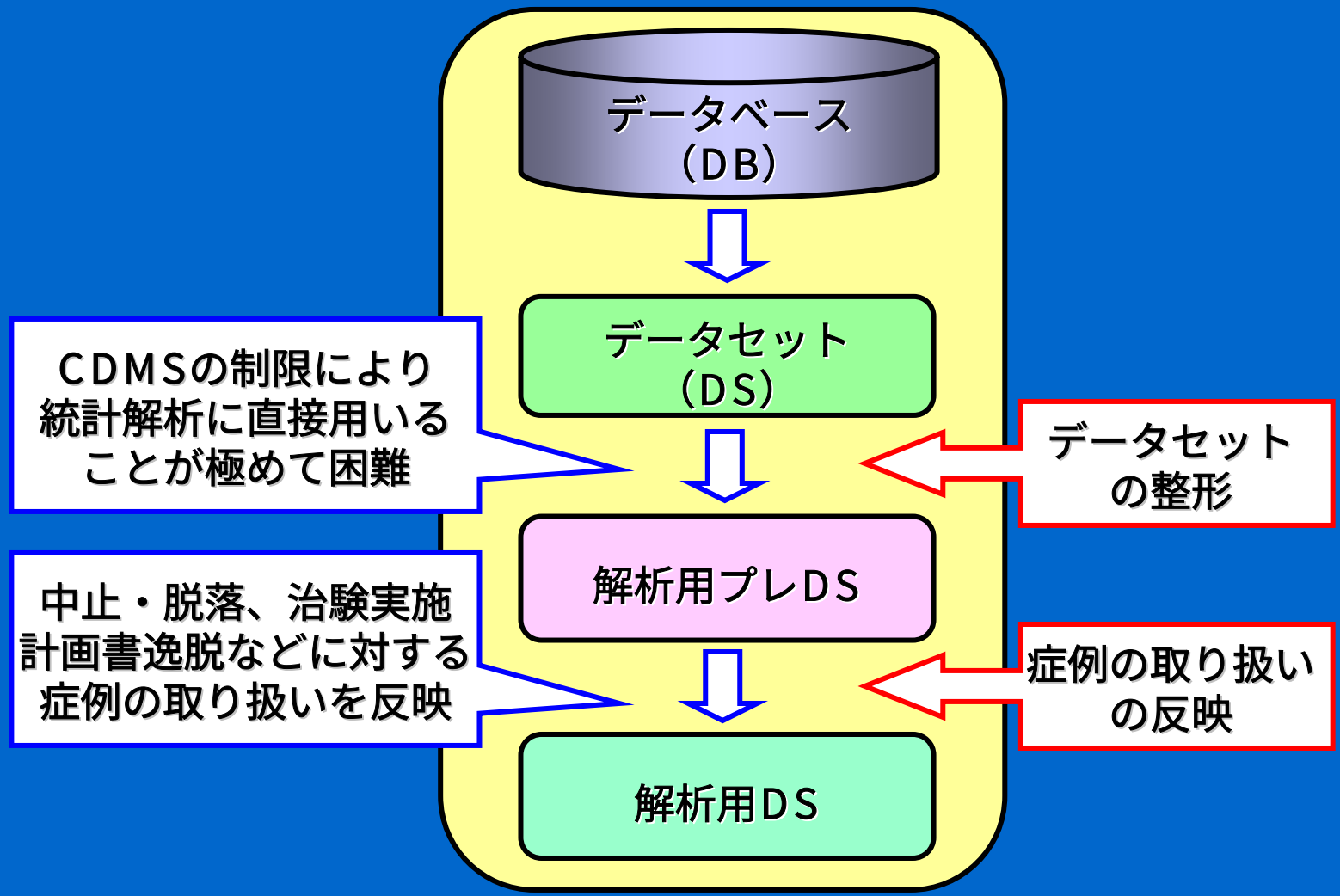
ダブルプログラミングを基礎
とした品質管理手法を評価する

本報告

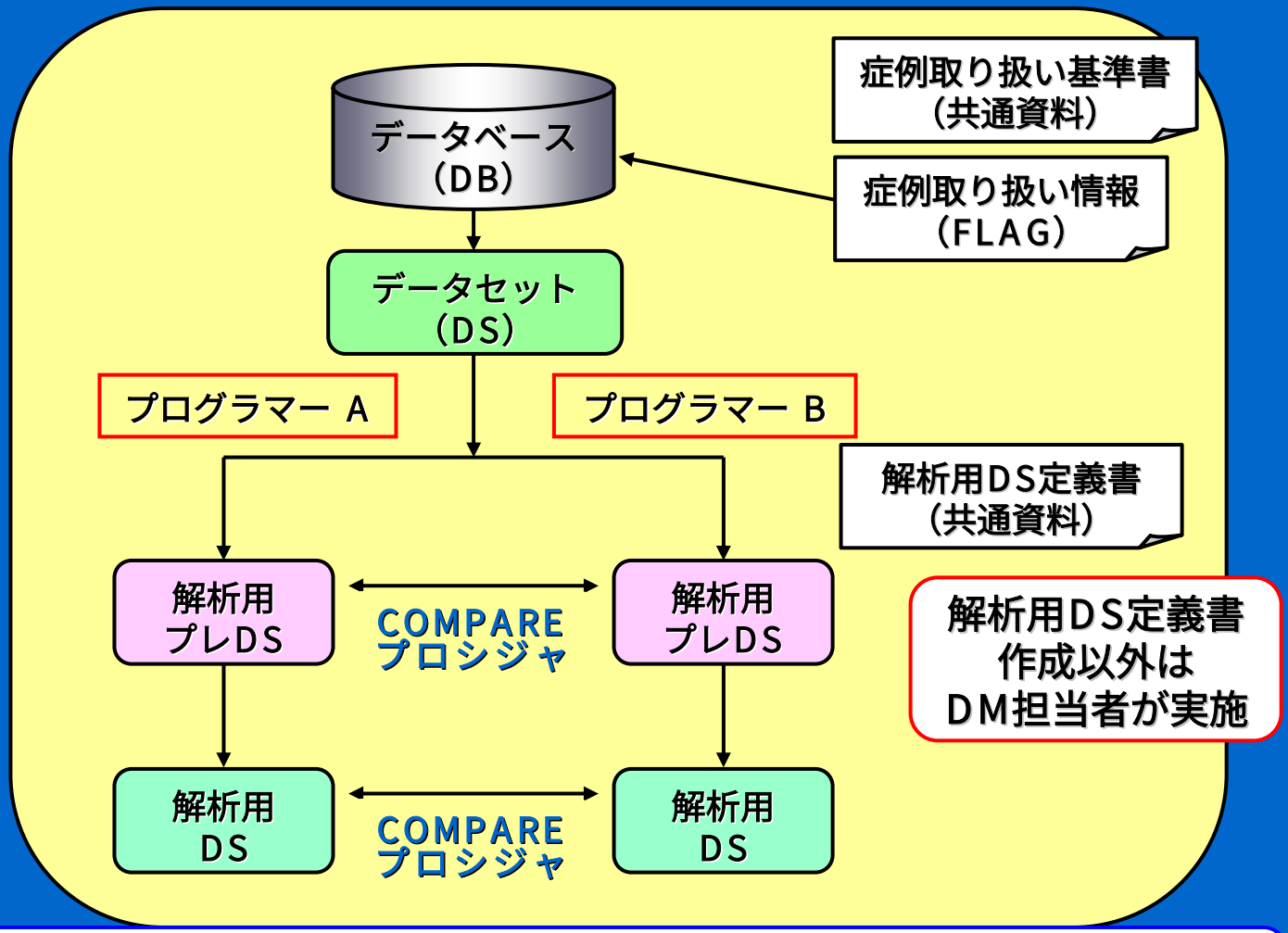
- ダブルプログラミングによる品質管理の手順
- ダブルプログラミングによる品質管理の実例
- ダブルプログラミングによる品質管理の評価
  - 目視による追跡
- コストの予測

ダブルプログラミングによる品質管理の手順

解析用ＤＳの作成手順


データベース
（DB）
データセット
（DS）
解析用プレDS
解析用DS
CDMSの制限により
統計解析に直接用いる
ことが極めて困難
中止・脱落、治験実施
計画書逸脱などに対する
症例の取り扱いを反映
データセット
の整形
症例の取り扱い
の反映

解析用ＤＳの作成手順



解析用プレＤＳ・解析用ＤＳ、各プログラム、各比較結果は
ＣＤＭＳの文書管理機能を用いて管理

COMPAREプロシジャの要約リスト

SAS システム

COMPARE プロシジャ
DMDATA.DRUG_STA と STDATA.DRUG_STA の比較
(METHOD=RELATIVE(2.22E-09), CRITERION=0.00001)

データセットの要約レポート

データ	作成日	更新日	VAR 数	OBS 数
DMDATA.DRUG_STA	26JAN04:10:04:29	26JAN04:10:04:29	16	19
STDATA.DRUG_STA	28JAN04:11:42:53	28JAN04:11:42:53	16	19

変数の要約レポート

共通変数の数 : 16.
ID 変数の数 : 2.

オブザベーションの要約レポート

OBS	基準	比較	ID
最初の OBS	1	1	CASENO=XXXX d__reno=1
最後の OBS	19	19	CASENO=XXXX d__reno=8

共通のオブザベーションの数 : 19.
DMDATA.DRUG_STA から読み込んだオブザベーションの数 (合計) : 19.
STDATA.DRUG_STA から読み込んだオブザベーションの数 (合計) : 19.

比較変数のうちどれかで等しくないオブザベーションの数 : 0.
すべての比較変数が同等なオブザベーションの数 : 19.

NOTE: 不等な値はありません。 比較した変数はすべて同等でした。

品質管理の記録になる

ダブルプログラミングによる品質管理の実例

検討に用いた試験と解析用プレＤＳの作成

- **検討に用いた試験**
  - １５症例の二重盲検ランダム化比較試験

- **解析用プレＤＳ作成**
  - プログラムの作成時間（含修正）　：３０時間×２
  - プログラムの長さ　：約１５００行
  - 結果の一致までの比較回数　：４回
  - 解析用プレＤＳのサイズ　：２４２obs
    × ４０６変数

解析用ＤＳの作成

- 解析用ＤＳの作成
  - 用意した症例の取り扱いの種類　：１７種類
  - 実際に発生した症例の取り扱い　：３種類
  - プログラムの作成時間（含修正）　：５時間×２
  - プログラムの長さ　：約２００行
約６００行
  - 結果の一致までの比較回数　：２回
  - 解析用ＤＳのサイズ　：２２５obs
×　４０６変数

検出されたエラーの実例

１．データのマージ
２．データの整形
３．ラベルの付加

実例１．　データのマージ

- 原因
  - 一方のプログラマーは測定時点の全てに初期値をマージ
  - 他方のプログラマーは初期値となる時点にのみ初期値をマージ
- 対策
  - 初期値をどの時点に作成するかを共通資料である解析用ＤＳ定義書に定義

変化量（率）を求めるには初期値が必要

解析用DS1

CASENO	TIME	VALUE	PRE_VALUE
101	1	18.5	18.5
101	2	18.7	・
101	3	19.3	・
101	4	17.6	・
102	1	20.3	20.3
102	2	22.1	・
102	3	21.4	・
102	4	21.8	・

解析用DS2

CASENO	TIME	VALUE	PRE_VALUE
101	1	18.5	18.5
101	2	18.7	18.5
101	3	19.3	18.5
101	4	17.6	18.5
102	1	20.3	20.3
102	2	22.1	20.3
102	3	21.4	20.3
102	4	21.8	20.3

実例２．データの整形（空白の除去）

- 原因
  - 一方のプログラマーはcompress関数を使用
  - 他方のプログラマーはcompress関数を使用せず

空白を含むデータは意図しない結果を生じることがある

変数値の比較結果

CASENO	d__reno	病歴名 基準値 ANAMNAME	比較値 ANAMNAME
101	1	花粉症2000年2月頃か	花粉症　2000年2月頃か
102	2	1993年花粉症(完治)	1993年　花粉症(完治)

薬剤名と投与量の間の空白がcompress関数により取り除かれ、本来の投与量が不明瞭になった

変数値の比較結果

	d__reno	薬剤名 基準値 DRUGNAME	比較値 DRUGNAME
101	1	ソリタT3200ml	ソリタT3　200ml
102	2	ソリタT3500ml	ソリタT3　500ml

実例２．　データの整形（空白の除去）

- 対策
  - trim関数とleft関数を組み合わせたプログラムを使用

```
%macro  label (taname, val1, val2) ;
      data &taname ;
            set &taname ;
            &val1.v = compress(&val1) ;
            drop &val1 ;
            label &val1.v = &val2 ;
            rename &val1.v = &val1 ;
      run ;
%mend ;
```

&val1.v = trim(left(&val1)) ;

実例３．　ラベルの付加

- 原因
  - 一方のプログラマーは全角の括弧を使用
    - » ラベルをタイプ
  - 他方のプログラマーは半角の括弧を使用
    - » ラベルをコピー&ペースト
- 対策
  - プログラマーは解析用ＤＳ定義書に記載された項目名をプログラムエディタにコピー&ペースト

ＤＢの項目名をラベルに自動的に取得するのが理想

ＣＤＭＳの制限により項目名を出力できない

属性が違う共通変数のリスト

Variable	Dataset	Type	Length	Label
AESYMPEXIST	DMDATA.BGRD_STA	Num	8	有害事象（自他覚）の有無
	STDATA.BGRD_STA	Num	8	有害事象(自他覚)の有無

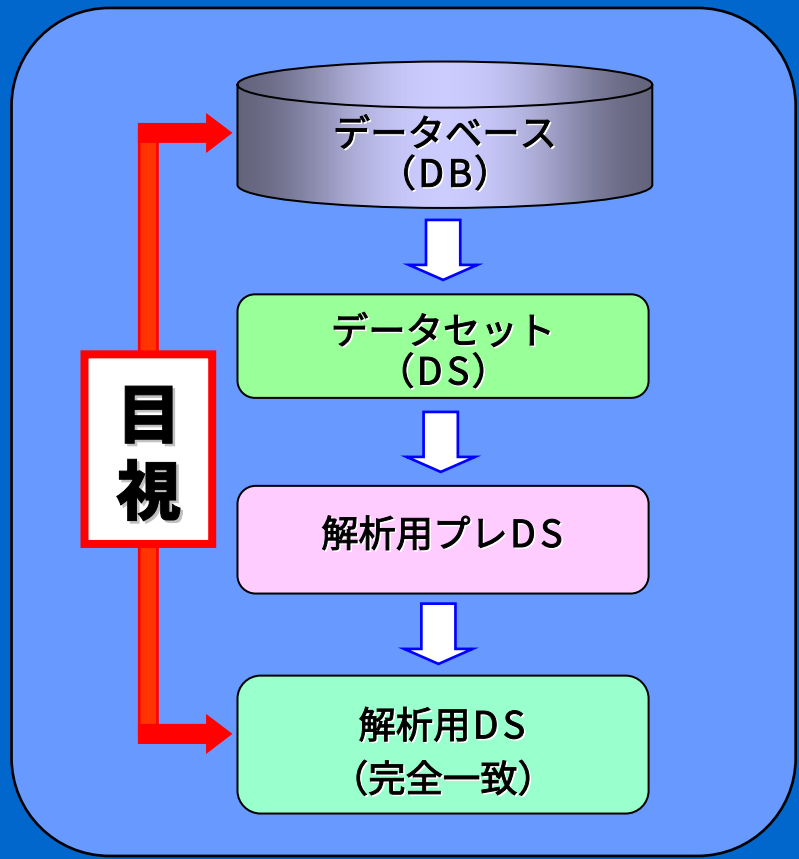
ダブルプログラミングによる品質管理の評価

目視による追跡

結果の比較を通過したエラーの検出

- 解析用DSのサイズ　：225obs × 406変数
- 作業人数　：2人
- 作業時間　：15時間
- 目視により発見されたエラー　：0件

３種類の症例の取り扱いはすべて適切に反映できていた



コストの予測

- 「シングルプログラミング ＋ 目視」
- 「ダブルプログラミング　＋ COMPAREプロシジャ」

シングルプログラミング
＋ 目視

プログラムの作成時間
　　解析用プレDS　：30時間
　　解析用DS　　：5時間
作成人数　　：1人（シングルプログラミング）
目視作業人数　　：2人
比較時間（解析用DS）
　　：15時間（目視）←サイズ（225obs × 406変数）
比較時間（解析用プレDS）
　　：16時間（目視）←サイズ（242obs × 406変数）

（30時間 ＋ 5時間 ±α）× 1人 ＋（16時間 ＋ 15時間 ）× 2人 × 2

＝ 約160人時間

エラーの検出に1回
エラーの修正に1回

ダブルプログラミング ＋ COMPAREプロシジャ

「ダブルプログラミング ＋ COMPAREプロシジャ」による方法

（30時間 ＋ 5時間）× ２人

＋（10秒 × 比較回数計6回）　＝ 約70人時間

「シングルプログラミング ＋ 目視」による方法

（30時間 ＋ ５時間 ±α）× 1人

＋（16時間 ＋ 15時間 ）× 2人 × 2

＝ 約160人時間

症例数が150例（15例の10倍）では、

（30時間 ＋ ５時間 ±α）× 1人

＋（16時間 ＋ 15時間 ）× 2人 × 2 ×10倍

＝ 約1300人時間

COPMPAREプロシジャ
完全一致に必要な比較回数
は症例数に強く依存しない

結果と結論

- 品質管理のためのトータルコストが小さい
  - 結果の比較コストが極めて小さい
  - ２重のプログラミングコストを吸収可能
  - 品質管理記録を作成する時間　０
- ダブルプログラミングによる品質管理は解析用ＤＳに対する効率的な品質管理である

最後に

- 解析用ＤＳはＤＭ担当者が作成すべき
  - ＤＭ担当者はデータを扱うプロフェッショナル
- 解析担当者は解析用ＤＳの受け入れ確認を行うべき
  - 異なる視点によるチェックは有用

謝辞

この論文を作成するにあたり業務量等様々な面で御配慮いただいた臨床解析部　吉田純朗　部長、データマネージメント課　大久保正人　課長に心より御礼申し上げます。
また、データの目視確認の補助や論文内容についての御助言を頂いたデータマネージメント課の皆様に感謝申し上げます。

参考文献

1）菅波秀規, 益田隆史. ダブルプログラミングによる統計解析の品質管理. 第19回. SUGI-J. 2000

2）菅波秀規. SASにおける統計解析バリデーションと解析計画書、報告書作成. 第7章. 臨床試験データ解析におけるダブルアナリシスの活用. 2003；113～142. 技術情報協会

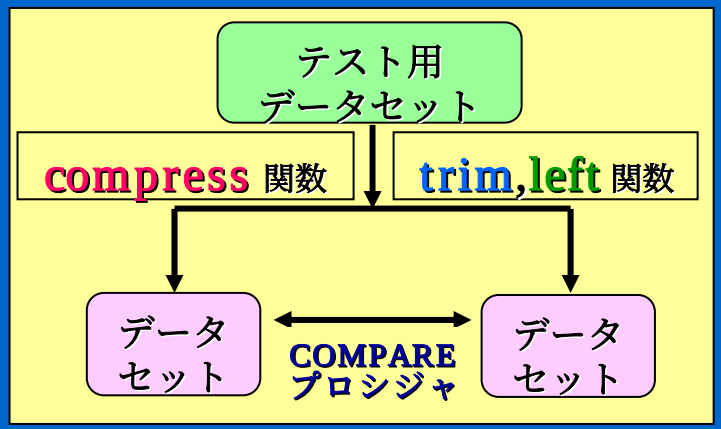
3）Ron Cody. Cody's Data Cleaning Techniques Using SAS Software. 1999；137～152. SAS Institute

バックアップスライド

参考　trim関数とleft関数使用時の結果

- データの先頭とデータ内に空白を入れたテスト用のDSを作成
- trim関数とleft関数の使用によりデータの前後の空白のみを取り除くことができた

	CASENO	COMMENT
1	101	XXXXX3 30mg



compress

変数値の比較結果

CASENO	基準値 COMMENT	比較値 COMMENT
101	XXXXX330mg	XXXXX3 30mg

trim,left

参考　ＤＭ部門が解析用ＤＳを作成している理由

- 統計解析担当者が症例の取り扱いを行うのは望ましくない
- ＤＭ担当者はＣＤＭＳに詳しい
  - ＤＢやＤＳの構造はＣＤＭＳに依存
- ＤＭ担当者は症例の性質やイレギュラーデータに詳しい
  - 症例報告書を詳細にレビューしている

参考

ＤＭ部門、統計解析部門による解析用ＤＳの確認作業

症例の取り扱い
情報の量や複雑さ

プログラマーの
能力の違い

系統的なエラー

検出されないエラーが
発生する可能性

ＤＭ担当者

解析用ＤＳ作成作業確認

統計解析担当者

解析用ＤＳ受け入れ確認

参考　ＤＭ部門での解析用ＤＳの確認作業

- 解析用ＤＳが適切なプロセスで作成されているかを確認
  - 最新の解析用ＤＳでないもので比較する危険性を回避

ＤＳの作成ログ(ＤＢ→ＤＳ：実行日時・正常終了)
解析用ＤＳ作成日 >= 解析用プレＤＳ作成日 >= ＤＳ作成日
解析対象集団のサイズ（ＡＲＳ >= ＦＡＳ >= ＰＰＳ） ＊ＡＲＳ: All Randomised Subjects
解析用ＤＳ作成用のプログラムの保存記録
ＣＯＭＰＡＲＥプロシジャ実行後の比較結果の保存記録
解析用ＤＳ(移送ファイル)の保存記録

ＤＢから最新のＤＳを抽出しているかを確認

最新のＤＳを用いて解析用ＤＳを作成しているかを確認

ＣＤＭＳの文書管理機能を用いて管理

参考　統計解析部門での解析用ＤＳの確認作業

- 受け入れ確認の内容

比較結果(要約リスト)の確認
解析対象集団の確認
テーブルのキー構造の確認
変数の属性とＬＡＢＥＬの確認
変数のレンジチェック
変数の欠測データ数の確認