統計解析結果のバリデーションを考える

大阪大学医学部附属病院
増村　一穂

目次

- 背景
- 解析結果の誤りはどこで発生するのか
- 工程毎のバリデーション
- まとめ

背景

N		120
性別	男性(1)	61(50.8%)
	女性(2)	59(49.2%)
年齢	50歳以上(1)	75(62.5%)
	50歳未満(2)	45(37.5%)

上記の解析結果を見て何か気づくことはありますでしょうか

背景

N		120
性別	男性(1)	59(49.2%)
	女性(2)	61(50.8%)
年齢	50歳以上(1)	75(62.5%)
	50歳未満(2)	45(37.5%)

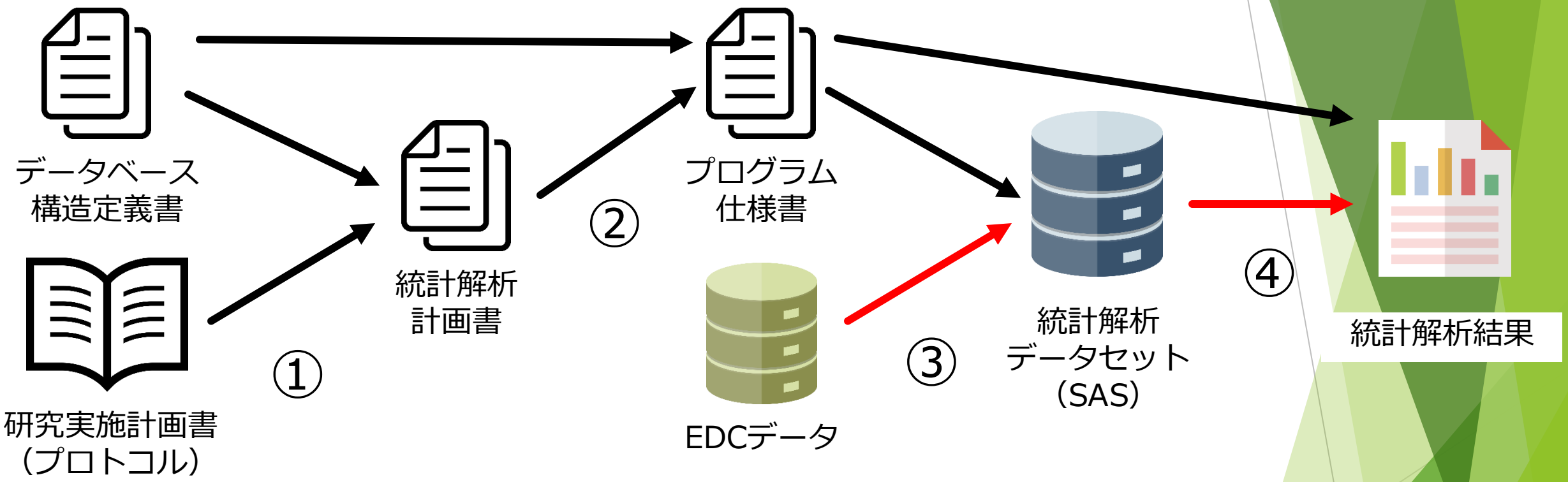
男性と女性の数が逆になっています

男性：１、女性２で集計すべき結果を
男性２、女性１で集計したことによるプログラム誤りが原因

解析結果の誤りは気づきにくい

正確な統計解析結果にはバリデーション（validation：確認）が必須

解析結果の誤りはどこで発生するのか



①：データベース構造定義書＋研究実施計画書 ⇒ 統計解析計画書
②：統計解析計画書 ⇒ プログラム仕様書
③：EDCデータ＋プログラム仕様書 ⇒ 統計解析データセット
④：統計解析データセット＋プログラム仕様書 ⇒ 統計解析結果

①～④、どこでも統計解析結果誤りの原因は発生する可能性がある

解析結果作成工程①

赤文字は突合
青文字は情報元が未記載

データベース構造定義書

登録情報	登録日	SUBJDAT	YYYY/MM/DD
血液検査	Visit	VISIT	Week1／Week2／Week3
	検査項目A	ITEMA	はい／いいえ
	検査値B	ITEMB	xx.xx
	検査値C	ITEMC	xx.xx
薬物療法	投与開始日	DRUGSTDAT	YYYY/MM/DD
	投与終了日	DRUGENDAT	YYYY/MM/DD
薬物療法後治療	投与開始日	ADRUGSTDAT	YYYY/MM/DD
	投与終了日	ADRUGENDAT	YYYY/MM/DD
生存情報	死亡の有無	DSYN	あり／なし
	死亡日	DSDAT	YYYY/MM/DD

研究実施計画書

解析１

・血液検査において初週における検査値の統計量を算出する。また、指標Dは検査値BとCの和より算出する

解析２

・全生存期間（Overall Survival：OS）（薬物療法開始から死亡までの期間）に対する解析を実施する

統計解析計画書

解析１

Week１における以下の統計量を算出する。
定性評価は頻度、割合を算出し、定量評価は総数、平均、標準偏差、最大値、最小値を算出する。また、指標DはB+Cとし、一方が欠測の場合はDも欠測とする。

・定性評価：検査項目A
・定量評価：検査値B、検査値C、指標D

解析２

死亡をイベントとしたKaplan-Meier曲線を描く。
開始日を薬物療法における投与開始日、
イベント日を死亡日とする。イベント未発生の場合は２年（730日）で打ち切りとする

グラフの描画にあたっては以下の通り作成する。
縦軸：イベント発生割合
横軸：50日刻み

解析結果作成工程②

赤文字は突合
青文字は情報元が未記載

統計解析計画書

解析１
Week１における以下の統計量を算出する。定性評価は頻度、割合を算出し、定量評価は総数、平均、標準偏差、最大値、最小値を算出する。
また、指標DはB+Cとし、一方が欠測の場合はDも欠測とする。
・定性評価：検査項目A
・定量評価：検査値B、検査値C、指標D

解析２
死亡をイベントとしたKaplan-Meier曲線を描く。
開始日を薬物療法における投与開始日、
イベント日を死亡日とする。イベント未発生の場合は２年（730日）で打ち切りとする

グラフの描画にあたっては以下の通り作成する。
縦軸：イベント発生割合
横軸：50日刻み

プログラム仕様書

解析データセット

項目A	ITEM_A	検査項目A：ITEMA where=(Visit = "week1")
項目B	ITEM_B	検査値B：ITEMB where=(Visit = "week1")
項目C	ITEM_C	検査値C：ITEMC where=(Visit = "week1")
項目D	ITEM_D	ITEM_B + ITEM_C ただしどちらかが欠損の場合は欠損とする
イベント起点日	EVENT_STDAT	【薬物療法】投与開始日：DRUGSTDAT
イベント発生	EVENT_YN	死亡の有無：DSYN 1：あり　0：なし
イベント発生日	EVENT_DAT	死亡日：DSDAT 死亡無しの場合はEVENT_STDAT+730
イベント発生期間	EVENT_TERM	EVENT_DAT – EVENT_STDAT

統計解析結果

解析１
ITEM_A：はい、いいえの頻度割合
ITEM_B、ITEM_C、ITEM_D：統計量算出

解析２
```
proc lifetest data=【INDATA】
  plots=survival(atrisk=0 to 730 by 50);
  time EVENT_TERM * EVENT_YN(0);
run;
```

DB構造定義書＋プロトコル　⇒　統計解析計画書
統計解析計画書　⇒　プログラム仕様書

発生する可能性のあるミス	対策（バリデーション）
DB構造定義書＋プロトコル　⇒　統計解析計画書 統計解析計画書　⇒　プログラム仕様書 **上記資料間にて実施すべき内容が記載されていない、もしくは誤った記載となっている**	**記載内容のマッピング** 元資料から作成された資料の間において内容の突合せを行い、記載漏れ、記載誤りを確認する。 確認結果に対する有識者（データ関連担当者、担当医師など）のレビューも重要である
似たような名称のデータを間違えて使用する 【例】 生存時間解析にあたって起点日は「**薬物療法の投与開始日**」を使用するが、誤って似たような名称の「**薬物療法後治療の投与開始日**」を使用してしまった	
未確認の情報がある 【例】 イベント未発生の場合の打ち切り日 生存時間分析グラフ描画にあたっての情報	**情報元（打ち合わせ記録など）の明確化** 打ち合わせによる決定事項であれば別資料として添付しておくのが望ましい

解析結果作成工程③

プログラム仕様書：解析データセット

項目A	ITEM_A	検査項目A：ITEMA where=(Visit = "week1")
項目B	ITEM_B	検査値B：ITEMB where=(Visit = "week1")
項目C	ITEM_C	検査値C：ITEMC where=(Visit = "week1")
項目D	ITEM_D	ITEM_B + ITEM_C ただしどちらかが欠損の場合は欠損とする
イベント起点日	EVENT_STDAT	【薬物療法】投与開始日：DRUGSTDAT
イベント発生	EVENT_YN	死亡の有無：DSYN 1：あり　0：なし
イベント発生日	EVENT_DAT	死亡日：DSDAT 死亡無しの場合はEVENT_STDAT+730
イベント発生期間	EVENT_TERM	EVENT_DAT – EVENT_STDAT

プログラム仕様書：統計解析

解析１
ITEM_A：はい、いいえの頻度割合
ITEM_B、ITEM_C、ITEM_D：統計量算出

解析２

```
proc lifetest data=【INDATA】
  plots=survival(atrisk=0 to 730 by 50);
  time EVENT_TERM * EVENT_YN(0);
run;
```

解析データセット

SUBJID	ITEM_A	ITEM_B	ITEM_C	ITEM_D
OSK001	はい	20	30	50
OSK002	いいえ	12	26	38
OSK003	はい	9	12	21
OSK004	いいえ	6	5	11
OSK005	いいえ		22	
OSK006	いいえ	18	6	24
OSK007	はい	5	2	7

SUBJID	EVENT_STDAT	EVENT_YN	EVENT_DAT	EVENT_TERM
OSK001	2020/1/1	0	2021/12/31	730
OSK002	2020/1/1	1	2020/1/10	9
OSK003	2020/1/1	1	2020/1/19	18
OSK004	2020/1/1	0	2021/12/31	730
OSK005	2020/1/1	0	2021/12/31	730
OSK006	2020/1/1	1	2020/1/30	29
OSK007	2020/1/1	1	2020/2/20	50

プログラム仕様書　⇒　解析データセット

発生する可能性のあるミス	対策（バリデーション）
プログラムコードミス	**ダブルプログラミングの場合** main側、sub側、それぞれが作成したSASデータセットに対するCompareプロシージャの実行 **シングルプログラミングの場合** 無し・・・（解析結果にて対策）

Compareプロシージャの差分観点　（**この資料の末尾にそれぞれの観点結果を記載しています**）

観点	異なっていた場合	備考、対策
変数の順番	問題無し	データセットと仕様書の可読性が悪くなるので 仕様書通りに作成することを推奨する
余分な変数（設計書に無い変数、フラグ変数など）	NG	仕様書以外の変数は残さない
変数の属性（長さ、フォーマット）	NG	仕様書に属性に関する内容を明確に記載する
ソート順	NG	仕様書にソートキーを明確に記載する

解析結果作成工程④

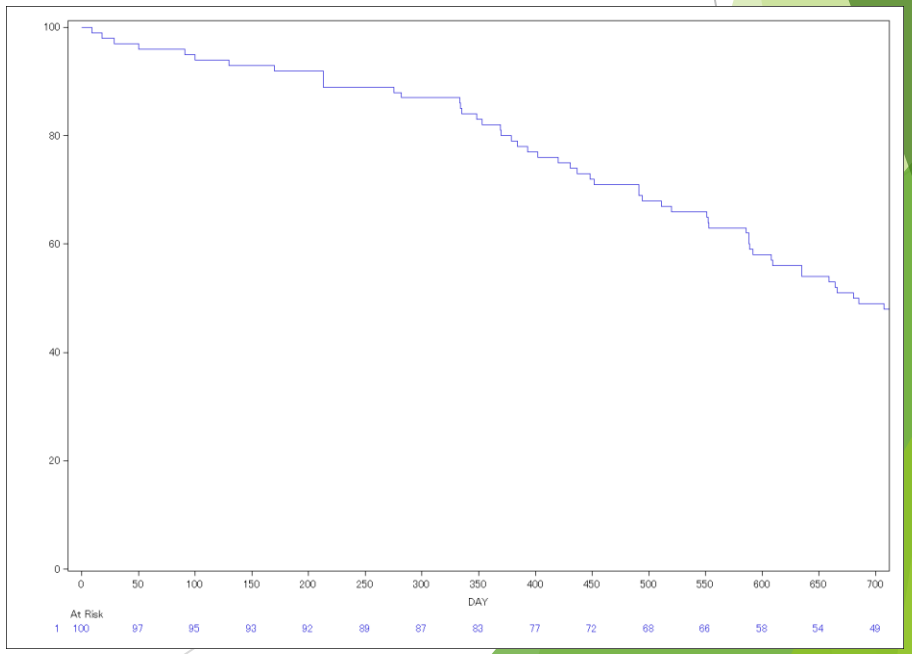
解析データセット

SUBJID	ITEM_A	ITEM_B	ITEM_C	ITEM_D
OSK001	はい	20	30	50
OSK002	いいえ	12	26	38
OSK003	はい	9	12	21
OSK004	いいえ	6	5	11
OSK005	いいえ		22	
OSK006	いいえ	18	6	24
OSK007	はい	5	2	7

SUBJID	EVENT_STDAT	EVENT_YN	EVENT_DAT	EVENT_TERM
OSK001	2020/1/1	0	2021/12/31	730
OSK002	2020/1/1	1	2020/1/10	9
OSK003	2020/1/1	1	2020/1/19	18
OSK004	2020/1/1	0	2021/12/31	730
OSK005	2020/1/1	0	2021/12/31	730
OSK006	2020/1/1	1	2020/1/30	29
OSK007	2020/1/1	1	2020/2/20	50

解析結果

N		7
検査項目A	はい	3(42.9%)
	いいえ	4(57.1%)
検査値B	N	6
	平均（標準偏差）	11.67(6.22)
	最小、最大	5 , 20
検査値C	N	7
	平均（標準偏差）	14.71(11.21)
	最小、最大	2 , 30
指標D	N	6
	平均（標準偏差）	25.17(16.31)
	最小、最大	7 , 50



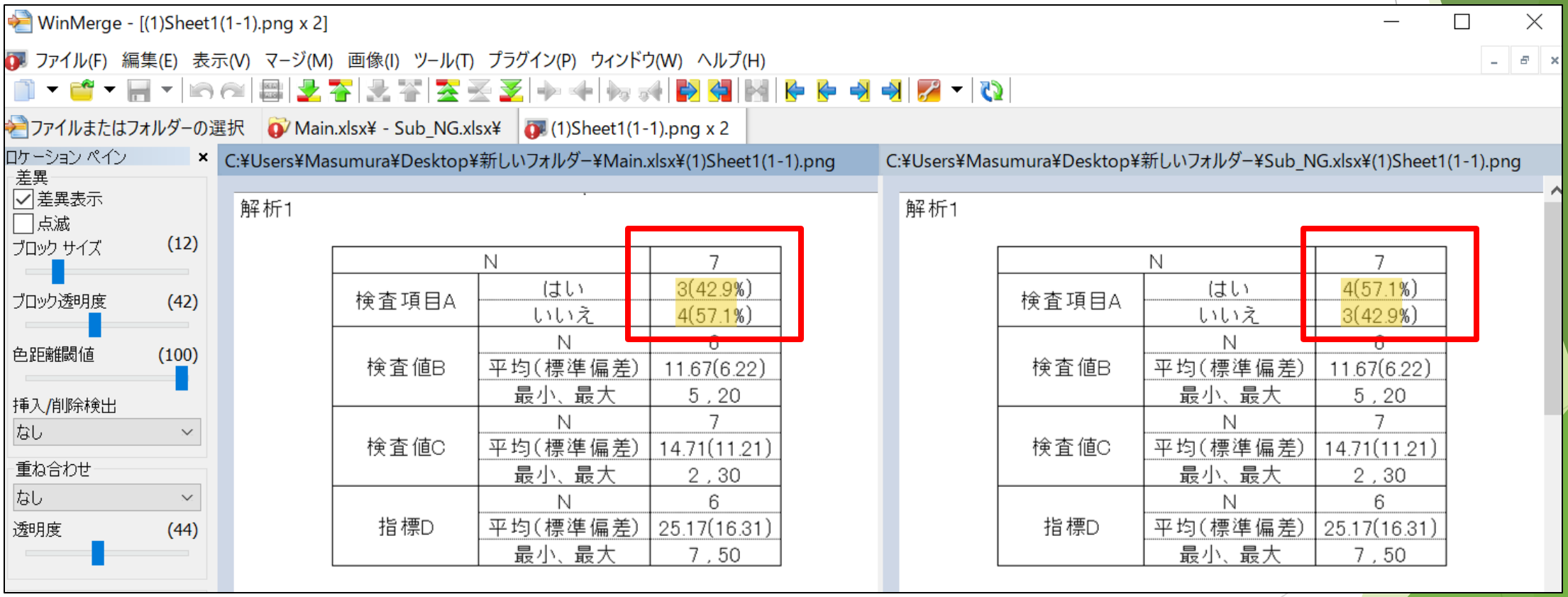
解析データセット　⇒　解析結果

発生する可能性のあるミス	対策（バリデーション）
プログラムコードミス	**ダブルプログラミングの場合** 解析結果（帳票）の比較（WinMergeなど） （資料の末尾にWinMergeによるExcel帳票比較の方法を記載しています）
	シングルプログラミングの場合 1人ダブルチェック（別方法で集計する）

様々な都合などもありますが、品質面から
基本的にダブルプログラミングを推奨します

WinMergeについて（Excel帳票、画像ファイルの比較）

表の差分を明確に出来ます

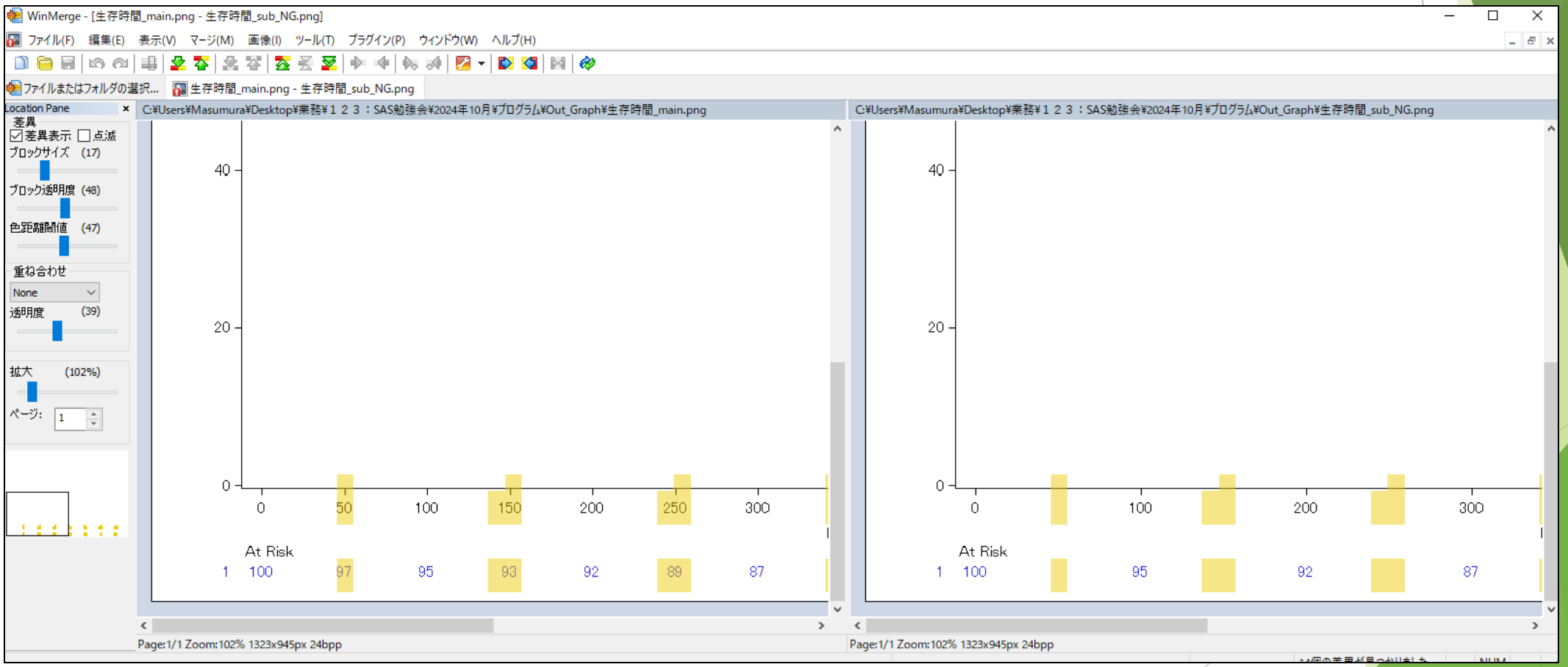


		7
検査項目A	はい	3(42.9%)
	いいえ	4(57.1%)
検査値B	N	6
	平均(標準偏差)	11.67(6.22)
	最小、最大	5 , 20
検査値C	N	7
	平均(標準偏差)	14.71(11.21)
	最小、最大	2 , 30
指標D	N	6
	平均(標準偏差)	25.17(16.31)
	最小、最大	7 , 50

		7
検査項目A	はい	4(57.1%)
	いいえ	3(42.9%)
検査値B	N	6
	平均(標準偏差)	11.67(6.22)
	最小、最大	5 , 20
検査値C	N	7
	平均(標準偏差)	14.71(11.21)
	最小、最大	2 , 30
指標D	N	6
	平均(標準偏差)	25.17(16.31)
	最小、最大	7 , 50

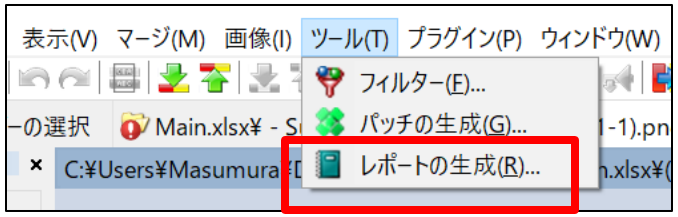
WinMergeについて（Excel帳票、画像ファイルの比較）

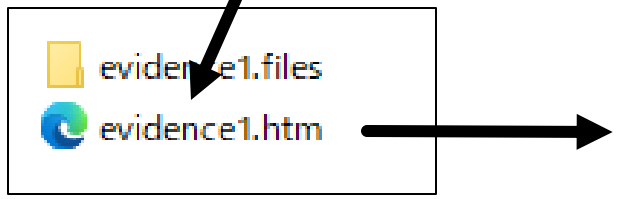
画像ファイル（図など）も比較できます

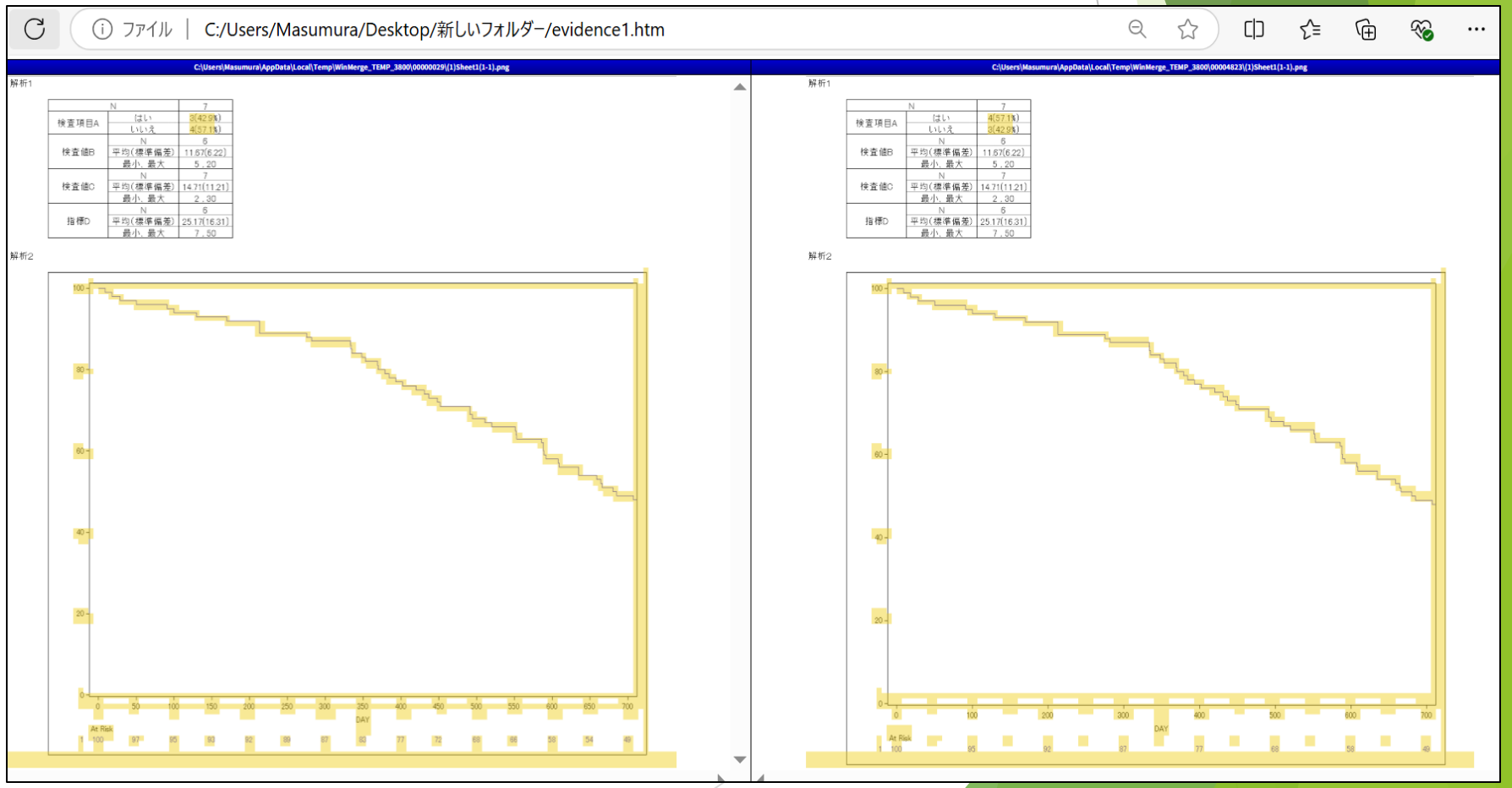


WinMergeについて（Excel帳票、画像ファイルの比較）

比較結果（差分証跡）ファイルも残せます







シングルプログラミングの場合のバリデーション

解析結果に対して別方法の集計を実施して確認する

【例】元データに対してExcel関数やフィルタを駆使して統計量や頻度を算出する

EDCデータ

統計解析結果

SUBJID	ITEM_A	ITEM_B	ITEM_C	ITEM_D
OSK001	はい	20	30	50
OSK002	いいえ	12	26	38
OSK003	はい	9	12	21
OSK004	いいえ	6	5	11
OSK005	いいえ		22	
OSK006	いいえ	18	6	24
OSK007	はい	5	2	7

N		7
検査項目A	はい	3(42.9%)
	いいえ	4(57.1%)
検査値B	N	6
	平均（標準偏差）	11.67(6.22)
	最小、最大	5 , 20
検査値C	N	7
	平均（標準偏差）	14.71(11.21)
	最小、最大	2 , 30
指標D	N	6
	平均（標準偏差）	25.17(16.31)
	最小、最大	7 , 50

SUM

AVERAGE　STDEV

MAX、MIN

デメリット

複雑な統計手法を用いた結果の確認は困難

まとめ

解析結果は一目では正しいかどうか判断できない。解析結果の確認のためにバリデーションは重要

解析結果の誤り原因は
DB構造定義書＋プロトコル　⇒　統計解析計画書　⇒　プログラム仕様書
⇒　統計解析データセット　⇒　統計解析結果
の**どの工程でも発生する可能性がある。**バリデーションは全工程に実施すべきである

DB構造定義書＋プロトコル　⇒　統計解析計画書　⇒　プログラム仕様書

記載内容のマッピング（資料突合）、情報元の明確化　が重要
ここで誤ると後工程でどれだけ丁寧に作業しても解析結果は誤ったものとなる

プログラム仕様書　⇒　統計解析データセット　⇒　統計解析結果

・Compareプロシージャ、WinMergeなど
・シングルプログラムは比較作業が困難となる、または品質面的に非推奨

ご清聴ありがとうございました