臨床研究における欠測データの統計解析：最新の動向と実践的な方法論について

野間 久史
情報・システム研究機構　統計数理研究所
2017年1月31日
国立がん研究センター 生物統計セミナー

URL: http://normanh.skr.jp/

Missing Data

- 一般的な統計学の教科書で解説される統計手法は、原則としてすべて「欠測はひとつもなく、完全なデータが観測されている」ことを前提としている
- しかしながら、ほとんどすべての調査・実験研究において、なんらかのデータの欠測は生じる
  - 長期間の追跡調査における脱落，追跡不能（Drop-out）
  - 質問紙調査における無回答，無記入
  - 計測機器の測定限界を超えるデータ

など

卵巣がんの予後因子研究

- Clark et al. (2001) は、英国 EdinburghのWestern General Hospital の診療データベースをもとに、1189人の卵巣がん患者のデータ（診断年：1984-1999年）から、予後因子の同定および予後予測モデルの構築に関する研究を行っている
  - がんの臨床研究では、患者の予後（症状の悪化や生存期間など）と関連する因子を同定することは、 症状の経過や治療の選択に重要な情報となるため、ひとつの大きな研究テーマとなっている

データベースの2次利用研究

- データベースを2次利用した「後ろ向き（retrospective）」に行われる研究では…
- あらかじめ、どのような研究を行い、統計解析にどのような変数の情報が必要になるか、明確に定められたもとでデータが集められたわけではないので、重要な変数にしばしば大きな割合で欠測が含まれる
  - データの精度の低さは、エビデンスレベルにも直結！！
- Clark et al. (2001) の予後因子研究でも、多くの変数に多数の欠測が含まれることが問題となっていた

欠測データの内訳（n=1189）

Variable	Name	Type	Levels	% missing
Albumin (*outcome variable*)	alb	Continuous (rounded)	30	33.0
Grade of tumour	grade	Ordinal	3	11.6
Residual disease	resdis	Ordinal	3	6.8
Performance status	ps	Ordinal	4	42.7
Presence/absence of ascites	ascites	Binary	2	5.4
Age (exact years)	age	Continuous	–	0.0
FIGO stage	figo	Nominal	4	1.8
Histology	histol	Nominal	7	0.0
Chemotherapy regimen	ctype	Nominal	3	0.0
Surgery (yes/no)	surg	Binary	2	0.0
CA125 (a cancer antigen)	ca125	Continuous	–	36.7
Alkaline phosphatase	alp	Continuous	–	33.1
All variables	–	–	–	70.1

Table 2: Variables and their missingness in the ovarian cancer dataset.

Royston and White (2011)

生存時間解析の回帰モデル

- Coxの比例ハザード回帰モデル（Cox, 1972）

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

  - ハザード（hazard; 瞬間のイベント発生率）について、結果変数（死亡までの時間）と共変量間の関数関係を規定した回帰モデル
  - 時間に関するコンポネント $\lambda_0(t)$ に特定の関数形を仮定しなくても、$\beta_1,\ldots,\beta_p$についての推測を行うことができる、セミパラメトリックモデル（打ち切りがある場合にも、推定可能）
  - $\beta_1,\ldots,\beta_p$ は相対リスク（対数ハザード比）と解釈することがで

Complete-Case Analysis

- 統計解析に用いるモデル（e.g., Cox回帰モデル）において、少なくとも1つの変数が欠測している対象者を、単純に、解析対象集団から除外する
- 残された対象者は、すべての変数が測定されているので、形式的に、完全データのデータセットができ上がる
- 通常の完全データに対する解析手法（最小二乗法，最尤法など）を適用することができる
- 長らく、多くの医学論文で、暗黙のうちに、慣用されてきた方法

Complete-Case Analysis

- 最も単純なアプローチとして、「結果変数と対象となる共変量のうち、少なくとも1つの変数が欠測している患者を除外した解析」を行うことを考えよう
- 238人（19.8%) の患者は、4つ以上の変数が欠測していた
- 少なくとも1つの変数が欠測している患者に至っては、831人（69.9%) にまで及び、Complete-Case Analysis を行う際のサンプルサイズは、1189人から358人（30.1%）にまで減少してしまう！！

欠測による問題①：バイアス

- 欠測が何の理由もなく「完全にランダムに」起こってくれるのであれば (MCAR) 、欠測を無視して解析を行っても、妥当な推定・検定を行うことができる
- しかしながら、一般的な医学研究で欠測を起こす対象者は「欠測を起こすなんらかの理由」がある
- 例えば、臨床試験で、「試験治療を受けていたが、症状が悪化したため」「有害事象が起こったため」という理由で脱落を起こすケースは、ランダムな欠測ではない
- 仮に、有意な関連が認められたとしても、それが欠測パターンの違いによるものであったら…？？

事例：Lurasidoneの第2相試験

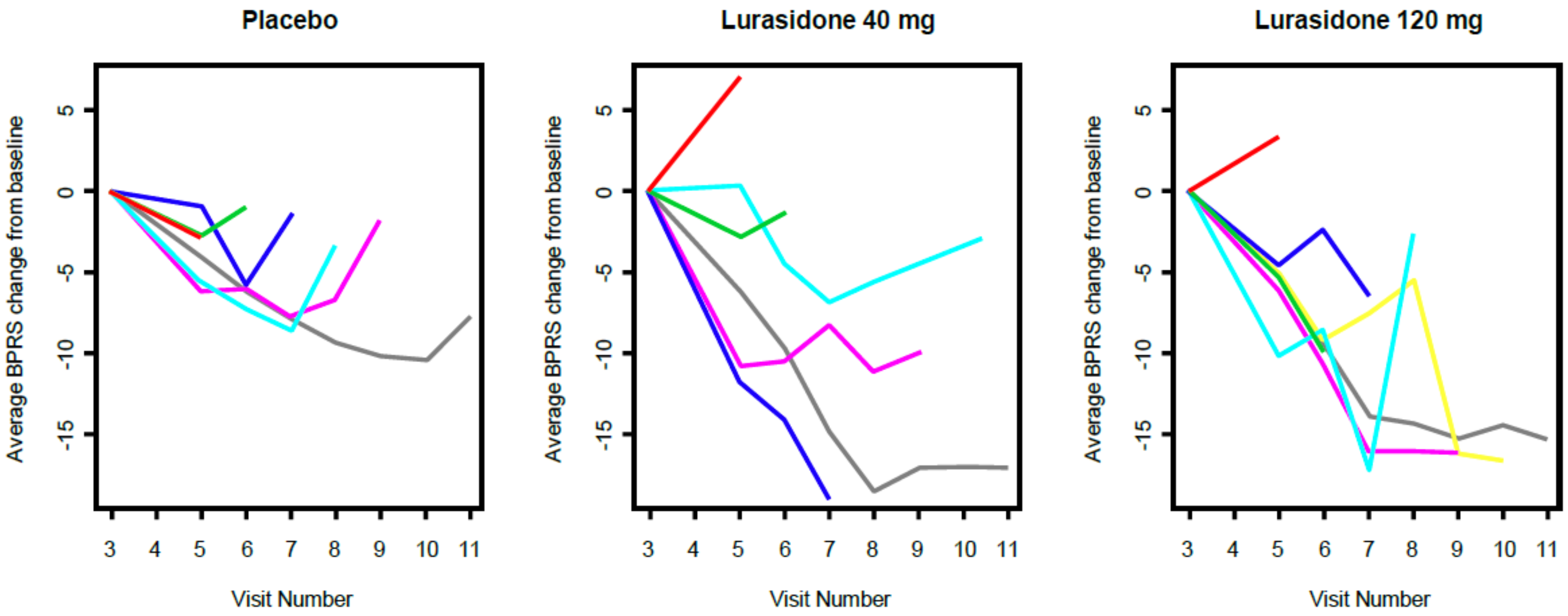


図 1　BPRS total score response profiles by treatment group

（審査官による解析．Statistical Review（s），Figure 2 より改変）

日本製薬工業協会 (2014)

脱落を起こした参加者を、脱落時点ごとにグループ分けし、グループごとのBPRSの時点ごとの平均値を、色分けしてプロットしたもの

欠測による問題②：推定精度・検出力の低下

- 欠測が起きると、統計的な推定・検定を行う上での情報量の損失が生じる
- せっかく多大な費用・労力をかけて行った研究でも、最終的な評価においては、欠測したデータの情報量の損失の分だけ、推定精度・検出力が低下してしまう
- 意味のある差があっても、欠測が多くなってしまうと、有意差が示せなくなってしまう可能性も

欠測による問題③：統計理論の想定外？？

- 線形回帰モデル

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

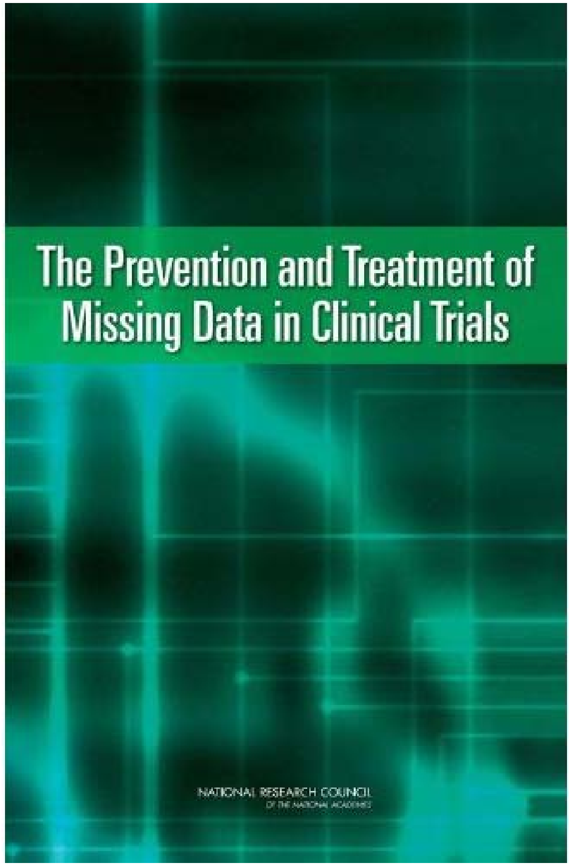
- 最小二乗法の二乗誤差関数

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})\}^2$$

- 結果変数，説明変数の組のうち、1つでも欠測があれば、そもそも通常の最小二乗法・最尤法における二乗誤差関数（尤度関数）が定義できない…？？

FDAの欠測データガイドライン

- 米国の規制当局（FDA）が医薬品開発における臨床試験での「欠測データの取り扱い」に関するガイドラインを作成することになった
- 米国 National Research Council (NRC) が、当該研究領域のエキスパートによる調査委員会を組織し（委員長：Roderick Little教授）、ガイドライン作成のための学術調査報告が行われた
- 解析手法だけではなく、欠測は、深刻なバイアスの原因となるため、あらかじめ最小限に抑えるための防止策と適切に取り扱うための包括的なRationaleが重要！！



http://www.nap.edu/catalog.php?record_id=12955

NRC Report: Recommendations

- 6. 試験のスポンサーは、欠測データが生じることによる潜在的な問題を明確に予想しておく必要がある。特に、プロトコルには、どの程度の欠測データが発生するかの見積もりや、欠測データの影響を最小限に留めるための試験計画、実施段階でのモニタリングの手順について明記した章も設けるべきである。
- 8.すべての試験のプロトコルは、欠測データを最小化することの重要性についての理解のもと、作成されるべきである。特に、過去の同様の試験の情報をもとにして、主要なアウトカムの情報の収集における最低限の達成目標を明確に定めるべきである。
- 9.試験のスポンサーは、欠測データの取り扱いに関する統計解析の方法について、プロトコルに明確に定めるべきである。また、それに関連する理論的な仮定についても、統計家だけではなく、臨床家が理解することができるように説明がされるべきである。

National Research Council (2010)

医学ジャーナルにおける動向

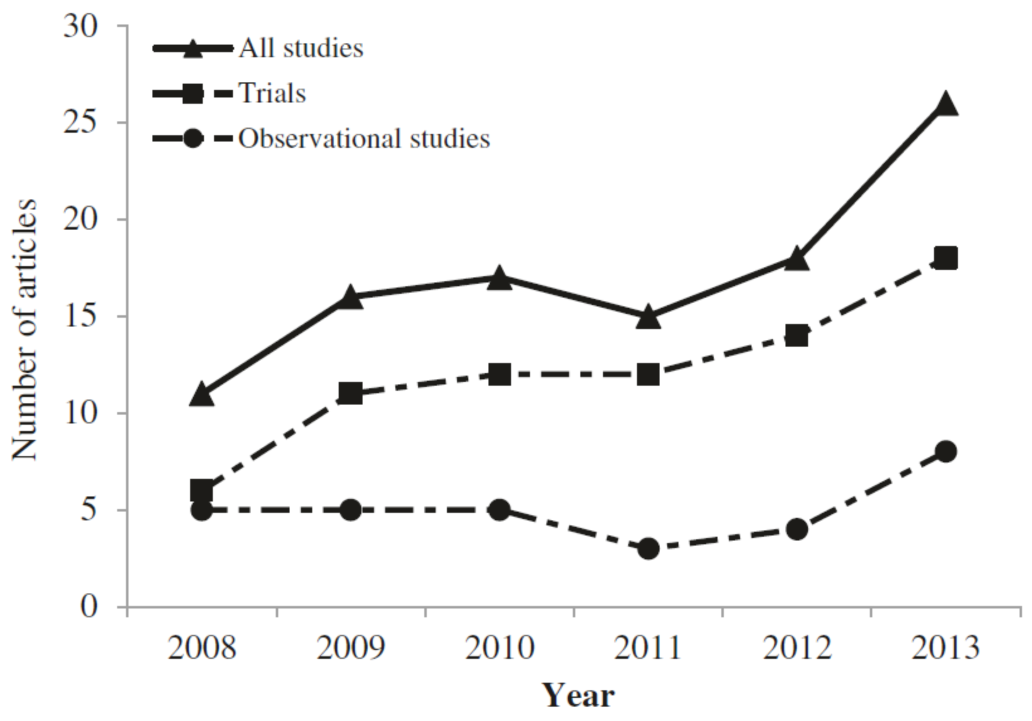


Figure 2 Number of articles in the Lancet and New England Journal of Medicine that used MI: overall and by study type.

- 多重代入法（multiple imputation; MI）に限った統計ではあるが、NRCによるRecommendationsが公表されて以降、一流医学ジャーナルの中でも、Complete-Case AnalysisのようなRoughな方法から、一定の理論的仮定のもとでの妥当な方法論へと、スタンダードとなる方法がシフトしつつある

Rezven et al. (2015)

NRC ReportによるRecommendationsから

- 一般的に、欠測データの統計解析は、高度な方法論・計算技法を用いたものが多く、また統計ソフトウェアにも標準的なモジュールとして実装されていないものが多かった
- 実質的には、学術誌の査読でも、不適切な解析（Complete-Case Analysisなど）が行われていても、見逃されてきた
- しかしながら、NRC ReportによるRecommendationsは、さまざまな学術コミュニティにおいて、これらの問題を再考する大きな契機となり、この数年ほどで、さまざまな領域で方法論に関する議論が行われるようになり、ソフトウェアの整備なども進められるようになった

欠測データの統計解析：Rationale

- まず、臨床試験において、すべての欠測データを統一的に扱う方法は存在しない
- 個々の試験のデザイン，測定値の特性などに応じて，必要な仮定・モデルは違う
- モデリングや推測の方法も広範に及び、いかなる状況においても万能な方法は存在しない
- Case-by-caseで、適切な解析手法を的確に選択する必要がある（それぞれの手法に対して正確な知識が必要！！）

Little et al. (2012)

欠測データに対する3つのシナリオ

- MCAR (Missing Completely At Random)
- MAR (Missing At Random)
- NMAR (Not Missing At Random)

Little et al. (2012)

MCAR

- すべての欠測は、完全にランダムに起こる（いかなる変数ともまったく無関係）
- Complete-Case Analysisで妥当な結論が得られる
  - 解析対象集団から、ランダムに一定の割合の対象者を除外することと同じ（除外後のデータも、母集団からの偏りのないランダムサンプルと見なすことができる）
  - ただし、検出力の低下は起こる

Complete Case Analysis

- NRC Reportでは、「MCARは、極めてあり得ない仮定である」と断じられている！！
- 臨床試験で、脱落や追跡不能が起こる場合、「何の理由もなくランダムに」という都合のよい仮定はまずあり得ない
- 脱落を起こす患者は、一般的に、脱落を起こすなんらかの理由がある（症状の悪化，副作用など）
- NRCレポートでは「Complete-Case Analysisは推奨しない」と明言されている

Little et al. (2012)

Clark and Altman (2003) の臨床研究でも

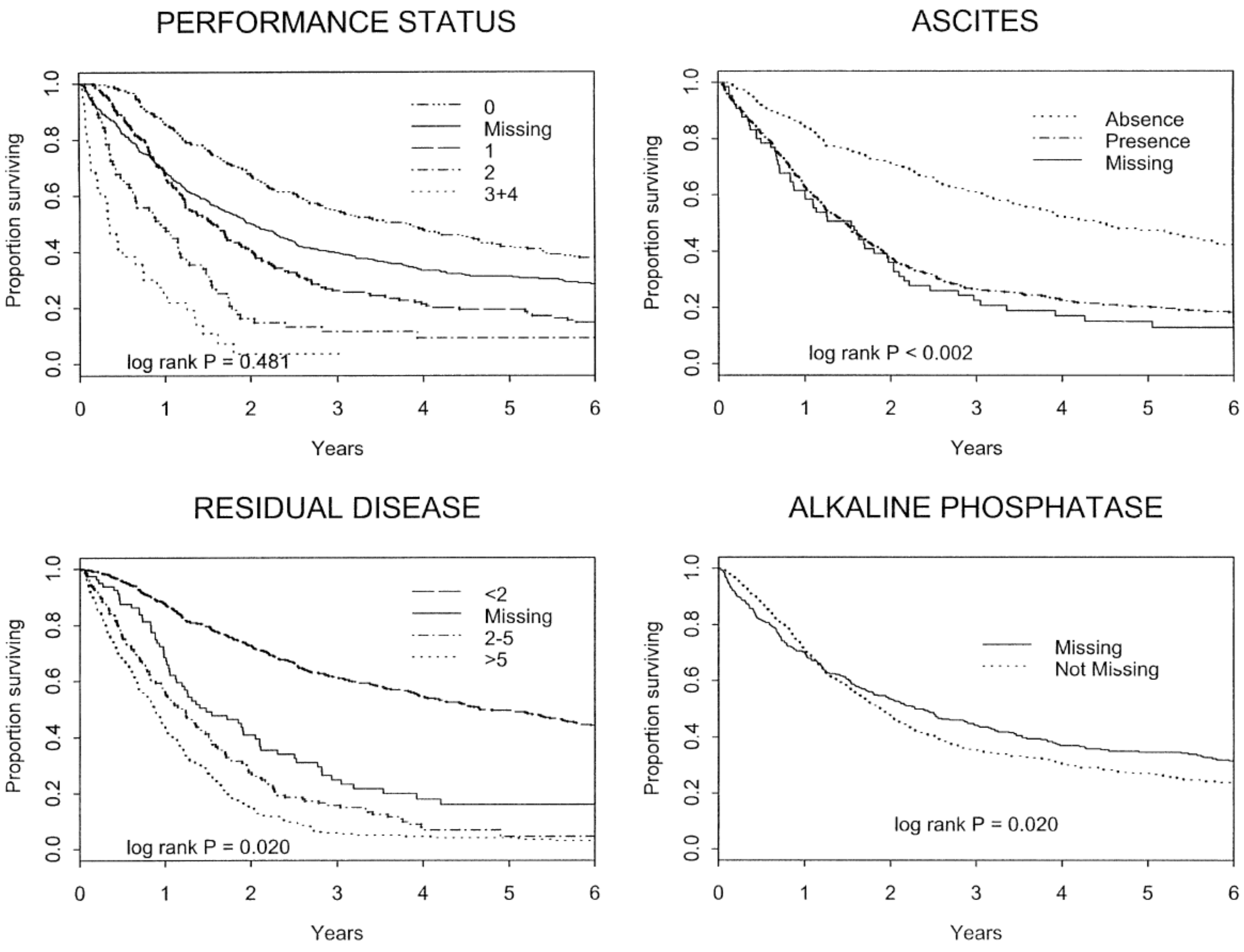


Fig. 1. Differences in survival of patients with and without recorded performance status, ascites, residual disease, and alkaline phosphatase, tested using log rank methods.

カテゴリごとのKaplan-Meier曲線と「欠測を起こした対象者」のKaplan-Meier曲線の比較

P値は、Observed vs. Missing の生存時間分布が一致するかどうかのログランク検定によるもの

有意差が出ている変数は、MCARではない！！

Clark and Altman (2003)

MAR

- 欠測のメカニズムは、観測されている変数ですべて完全に説明することができる
- 観測されたデータによって、（非ランダムな）欠測メカニズムを考慮した統計解析を行うことが可能
- NRC Reportでは、MARのもとで妥当性が保証される、以下の方法を Recommend している
  - 重み付き推定方程式による方法（IPW法）
  - モデルに基づく方法（多重代入法，直接尤度法，ベイズ法）

NMAR

- 欠測のメカニズムは、観測されている変数では完全に説明することができず、観測されていない変数にも影響される
- 感度解析をするしかない!!
  - パターン混合モデル
  - 選択モデル
- NNARのもとでの解析は、やや高度な手法が必要となり、まだ標準的な統計ソフトウェアでの整備も進められていない
- 医学ジャーナルの論文審査において、これらの感度解析の普及が進むのは、まだ当分先のことであると思われる

- ご関心がおありの方は、高井ら (2016)，日本製薬工業協会 (2016) などをご参照ください

多重代入法（Multiple Imputation）

- 現状の医学ジャーナルにおける、臨床研究・疫学研究の統計解析において、最も普及している方法
- データセット中の欠測値に対して、適当な代入値を埋め込むことによって、擬似的な完全データを作成し、これによって、すべての Available Data を用いたCox回帰などの解析を行うことができるようにする

Rubin (1987)

欠測値の代入（Imputation）とは？

X_1	X_2	X_3	X_4	X_5
0	-0.125	1.129	0.049	1.084
1	0.694	0.602	1.018	NA
0	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	0.870
1	0.327	-1.527	-1.459	NA
1	-0.243	-1.488	-1.449	-1.132
...	...	...	...	...

-1.240

字義通り、欠測値をなんらかの値で埋める操作のこと

1.084

単一代入法（Single Imputation）

- 欠測値を補完することで、見かけ上は「欠測のない完全データ」ができるので、通常のCox回帰分析を行うことができる
  - 観測されている変数についての情報をフルに活用することができ、推定精度・検出力もRecoveryできる
- 直観的には、「欠測してしまったデータ」を正確に予測して、それに近い値を代入することができれば、バイアスをなくす（正しく調整する）ことができそうである
  - 実際に、一定の条件下でバイアスをなくすことはできる
- しかしながら、単一の補完値に基づく統計解析の方法が、推奨されている文献はない

単一代入法の難点

- 代入後のデータセットに、通常のCox回帰を行って得られるP値と信頼区間は「すべてのデータが観測されたという前提のもの」となる
- 実際には、欠測した値は未知であり、それに近いと思われる値を予測して、置き換えているだけ
- 欠測値を100%正しく予測できるのであれば、妥当な結果に
- しかし、原理的に、未知の値を100%正しく予測することはできないため、欠測値の予測には必ず不確実性が生じる
- 最終的なP値，95%信頼区間が、その不確実性を無視していることになるため、統計的な誤差を過小評価することに

多重代入法（Multiple Imputation）

X_1	X_2	X_3	X_4	X_5
0	-0.125	1.129	0.049	1.084
1	0.694	0.602	1.018	NA
0	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	0.870
1	0.327	-1.527	-1.459	NA
1	-0.243	-1.488	-1.449	-1.132
…	…	…	…	…

-1.240
-0.898
⋮
-1.338

1.890
0.911
⋮
1.084

1つではなく、複数の補完値を用いる！！

なぜ、複数の補完値？？

- 補完値の予測の不確実性は、最終的なP値，信頼区間（＝回帰パラメータの推定量の分散）に反映される
- 補完値の予測の不確実性による、最終的な回帰パラメータの分散の増分は、複数の補完値ごとの解析結果のばらつきによって評価することができる
- 多重代入法は、この分散の評価におけるバイアスを適切に補正することによって、妥当なP値と信頼区間を与える方法となる
- 一方、単一代入法から得られるP値，信頼区間は誤り（Type-1 Error Rateを制御できない、誤差を過小評価する）

多重代入法のアルゴリズム①

- 欠測値に対して複数の補完値（M組）を生成
  - 補完値の生成方法はいろいろ（後ほど）
- 補完値を埋め込むことによってできる、M組の擬似的な完全データに対して解析を行い、推定値とその分散を求める
  - $\hat{\beta}_h, \hat{V}(\hat{\beta}_h), h = 1,2, \dots, M$

多重代入法のアルゴリズム②

- M 回の推定結果を統合（Rubinの統合公式）

$$\hat{\beta}_{\mathrm{MI}} = \frac{1}{M}\sum_{h=1}^{M}\hat{\beta}_h$$

$$\hat{\mathrm{V}}(\hat{\beta}_{\mathrm{MI}}) = \underbrace{\frac{1}{M}\sum_{h=1}^{M}\hat{\mathrm{V}}(\hat{\beta}_h)}_{\text{完全データの推定量の分散}} + \underbrace{(1+M^{-1})\frac{\sum_{h=1}^{M}(\hat{\beta}_h-\hat{\beta}_{\mathrm{MI}})(\hat{\beta}_h-\hat{\beta}_{\mathrm{MI}})^T}{M-1}}_{\text{欠測値の予測の不確実性によって生じる付加的なばらつきを表す項}}$$

Rubin (1987)

単一代入法と多重代入法

- 多重代入法で生成する補完値のデータとは？
  - 欠測したデータの背景にある確率分布（真の構造）から生成されるデータ（いわゆる乱数）
- 1点としての「正確な値」は予測できないが、観測されたデータから、欠測したデータが「どのような分布に従って得られるはずのものであったか」を推定することはできる
- その分布からの値を、複数シミュレーションして、その「分布の情報」を補完値として組み込む（というイメージ）

代表的な補完値の生成方法

- 連続変数：線形回帰モデル，予測平均マッチング
- カテゴリカル変数（2水準）：ロジスティック回帰モデル
- 順序を持たないカテゴリカル変数（3水準以上）：多項ロジスティック回帰モデル
- 順序を持つカテゴリカル変数（3水準以上）：多項ロジスティック回帰モデル，比例オッズモデル
- その他：マルコフ連鎖モンテカルロ法，ベイズ流（近似）ブートストラップ法　など

White et al. (2011)

線形回帰モデルによる代入値生成

- 欠測した変数を結果変数として、観測されているデータで、その分布を説明する回帰モデルを作る

$$X_1 = \gamma_1 + \gamma_2 x_2 + \gamma_3 x_3 + \cdots + \gamma_q x_q$$

- 最小二乗法によって推定された回帰式に、推定された誤差分散を上乗せさせて、X_1 が欠測した患者の X_1 の予測値を乱数によって発生させる
- この予測値を、代入値として用いるという方法

予測平均マッチング

- 回帰モデルによる方法で推定された回帰式と誤差分散には誤差が伴う（誤差の分布が正規分布から外れることも）
  - Plausibleでない値が生成されることも当然ある
- よりPlausibleな値に近づけるために、推定された回帰式から予測される「予測平均」に近い観測値が得られている他の対象者（例えば、近い順に5人）のデータから、補完値をランダムに選びとる（マッチングさせる）という方法
  - *Hot-deck Imputation*といわれる補完方法
- シミュレーションなどによる経験的な評価では、相対的に良好な性能を持つ方法であることが知られている

ロジスティック回帰モデルによる代入値生成

- 線形回帰モデルと同じく、欠測した変数を結果変数として、観測されているデータで、その分布を説明する回帰モデルを作る

$$\operatorname{logit} \Pr(X_1 = 1) = \gamma_1 + \gamma_2 x_2 + \gamma_3 x_3 + \cdots + \gamma_q x_q$$

- 最尤法によって推定された推定値 $\hat{\boldsymbol{\gamma}}$ とその漸近共分散行列の推定値 $\hat{\boldsymbol{V}}$ に対して、$MVN(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{V}})$ から乱数 $\boldsymbol{\gamma}^{*(1)}, \ldots, \boldsymbol{\gamma}^{*(M)}$を発生させ、これをplug-inした上記のロジスティック回帰の予測式から、欠測した X_1 の予測値（代入値）を生成する
- 多項ロジスティック回帰，比例オッズモデルも同様

単純な代入値の生成方法の問題点

- ここまで述べてきた、回帰モデルによる方法などでは、原則として、欠測を起こす変数が1つのみであることを想定している（SAS, Rなどに実装されているソフトウェアでも）
- しかし、一般的な調査・実験研究では、欠測を起こす変数が1つのみという性質のよい状況は、ほとんどあり得ない
- Clark et al. (2001) の卵巣がん研究でも、患者ごとにさまざまな組み合わせで複数の変数に欠測が起こっていた
- この場合、「欠測変数の予測のための回帰モデル」の説明変数が欠測することになり、欠測値の予測のためのモデルでも、また欠測に悩まされることになってしまう？？

複数の変数に欠測がある設定 (Example)

X_1	X_2	X_3	X_4	X_5
0	NA	1.129	0.049	1.084
1	0.694	NA	1.018	-1.240
NA	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	NA
NA	NA	-1.527	-1.459	1.084
1	-0.243	-1.488	NA	-1.132
…	…	…	…	…

連鎖方程式による多重代入法 (MICE)

- Multiple Imputation by Chained Equation (MICE)
- 複数の変数にまたがって、多重代入法による解析を行うための方法として開発された方法
- 基本的には、ここまで説明した多重代入法のアルゴリズムを組み合わせることによって実行することができる
- 個人ごとに、異なる組み合わせで複数の変数に欠測が起こっていても適用することができる
- 欠測を起こしている変数の型が異なっていて、異なる代入値の生成方法が必要とされる場合にも、適用することができる

White et al. (2011), Royston and White (2011)

MICEに必要となる仮定

- Fully Conditionally Specified (FCS) の仮定
  - $X_1, X_2, \ldots, X_p$ という変数の組は、すべての変数が、他の $(p-1)$ 個の変数によって、お互いの分布を完全に説明することができるという仮定
- X_1 の分布は、$X_2, X_3, \ldots, X_p$ で完全に説明できる
- X_2 の分布は、$X_1, X_3, \ldots, X_p$ で完全に説明できる
- …

MICEのアルゴリズム①

- FCSの仮定のもとで、$X_1, X_2, \dots, X_p$のそれぞれの変数についての補完値を順繰りに（連鎖的に）生成していく
- 例：X_1についての代入値を生成する場合
  - $X_2, X_3, \dots, X_p$　の欠測値には、1時点前の補完値を入れておき、その擬似的な完全データに対して、X_1の予測モデルを構築する
  - X_1　についての補完値を1組生成し（回帰モデルによる方法，予測平均マッチングなどで）、これをCurrentの値として更新する

Original Dataset (Incomplete Data)

X_1	X_2	X_3	X_4	X_5
0	NA	1.129	0.049	1.084
1	0.694	NA	1.018	-1.240
NA	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	NA
NA	NA	-1.527	-1.459	1.084
1	-0.243	-1.488	NA	-1.132
…	…	…	…	…

Initial Setting (Naïve Imputation)

X_1	X_2	X_3	X_4	X_5
0	0.694	1.129	0.049	1.084
1	0.694	-1.527	1.018	-1.240
1	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	-1.240
0	-0.809	-1.527	-1.459	1.084
1	-0.243	-1.488	0.049	-1.132
...	...	...	...	...

Cycle for X_1

X_1	X_2	X_3	X_4	X_5
0	0.694	1.129	0.049	1.084
1	0.694	-1.527	1.018	-1.240
NA	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	-1.240
NA	-0.809	-1.527	-1.459	1.084
1	-0.243	-1.488	0.049	-1.132
	…	…	…	…

X_1の予測モデルの構築

Cycle for X_2

X_1	X_2	X_3	X_4	X_5
0	NA	1.129	0.049	1.084
1	0.694	-1.527	1.018	-1.240
1	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	-1.240
0	NA	-1.527	-1.459	1.084
1	-0.243	-1.488	0.049	-1.132
...	...	...	...	...

X_2の予測モデルの構築

Cycle for X_3

X_1	X_2	X_3	X_4	X_5
0	0.521	1.129	0.049	1.084
1	0.694	NA	1.018	-1.240
1	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	-1.240
0	0.023	-1.527	-1.459	1.084
1	-0.243	-1.488	0.049	-1.132
...	...	...	...	...

MICEのアルゴリズム②

- 以上の連鎖的なアルゴリズムを $X_1, X_2, \ldots, X_p$ に順繰りに繰り返していき、M組の代入値の組を作成
  - Multiple Imputation by Chained Equation (MICE)
- 経験的に良好な性能を持つ方法であることも知られてきており、標準的なソフトウェアにも実装されてきている
  - SAS PROC MI: fcs statement
  - R, S-PLUS library: mice
  - Stata module: ICE

補足：初期値とBurn-in periodの設定

- MICE は、Roughな初期値（リサンプリングなどでデタラメに決めることが一般的）からスタートして、上記のアルゴリズムを何度も何度も繰り返すことによって、徐々に「適当な代入値が生成する定常分布からのサンプリング」に収束していくということを想定した方法
  - マルコフ連鎖モンテカルロ法をモデルとしている
- はじめのほうの数十回程度の代入値は、”burn-in period” として捨てた上で、以降のM組を多重代入法に用いる代入値とすることが一般的である

Case Study: 卵巣がんの臨床研究

- R, S-PLUS package: mice
  - もともとは、1990年代に S-PLUS のモジュールとして開発された
  - 現状の多重代入法のソフトウェアパッケージの中で、最も長い間、実践でも用いられてきた（多くの研究者により、バリデーションもされてきた）
- SAS, Stataでも同様の解析を実行することはできる
- 事例データ，プログラムは、以下のURLから入手できる
  - http://normanh.skr.jp/materials.html

R example: Complete-Case Analysis

```
ph1 <- coxph(Surv(t, d) ~ age + figo + grade + histol + ascites + ps + resdis + log_ca125 + log_alp, data=tgce)
summary(ph1)

Call:
coxph(formula = Surv(t, d) ~ age + figo + grade + histol + ascites +
    ps + resdis + log_ca125 + log_alp, data = tgce)

  n= 362, number of events= 248
   (827 observations deleted due to missingness)

                                    coef exp(coef)  se(coef)      z Pr(>|z|)
age                             0.017297  1.017447  0.006798  2.545 0.010943 *
figoII                          0.869917  2.386712  0.377383  2.305 0.021159 *
figoIII                         1.388945  4.010617  0.325320  4.269 1.96e-05 ***
figoIV                          1.762204  5.825262  0.372662  4.729 2.26e-06 ***
grademoderately differentiated  0.248357  1.281918  0.344870  0.720 0.471434
gradepoorly differentiated      0.118153  1.125416  0.325479  0.363 0.716596
histolAdenocarcinoma            0.428714  1.535282  0.724775  0.592 0.554176
histolEndometrioid             -0.039115  0.961640  0.315038 -0.124 0.901188
histolMesonephroid (clear cell) 0.535352  1.708049  0.274731  1.949 0.051338 .
histolMixed mesodermal          1.383109  3.987278  0.410101  3.373 0.000745 ***
histolMucinous                 -0.265081  0.767144  0.173456 -1.528 0.126455
histolUndifferentiated          0.356664  1.428555  0.443565  0.804 0.421349
ascitespresent                  0.298751  1.348174  0.164309  1.818 0.069029 .
ps1                             0.150572  1.162499  0.152834  0.985 0.324525
ps2                            -0.032588  0.967938  0.225161 -0.145 0.884924
ps3to4                          0.832012  2.297938  0.443744  1.875 0.060795 .
resdis2-5 cm                   -0.028585  0.971820  0.190122 -0.150 0.880488
resdis<2 cm                    -0.567408  0.566993  0.189456 -2.995 0.002745 **
log_ca125                       0.062361  1.064347  0.049242  1.266 0.205366
log_alp                         0.477335  1.611773  0.175360  2.722 0.006488 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ほとんどすべての統計ソフトは、デフォルトで
欠測データを除外し、Complete-Case Analysisを行う

R Example: MICE

```
library(mice)    # mice というパッケージを読み込む（install.package(“mice”) でインストールできる）

predmt1 <- (1 - diag(1, ncol(tgce)))
predmt1[c(1,11,12),] <- predmt1[,c(1,11,12)] <- 0
      # 代入値の生成モデルにおいて、それぞれの変数の予測モデルの説明変数の指定（行列形式）
```

```
imp.tgce <- mice(tgce, m=200, predictorMatrix=predmt1, seed=34871)
      # 補完値の生成方法については、次頁（ここではデフォルトの方法をそのまま採用；method で指定することができる）
      # m は、代入値の組の数（デフォルトでは5であるが、ここでは200としている）
      # predictorMatrixで、予測モデルに含める変数の組を指定している
      # seedは、乱数のシード
```

```
complete(imp.tgce, 5)   # 5番目のimputed datasetを出力
```

```
ph2 <- with(imp.tgce, coxph(Surv(t, d) ~ age + figo + grade + histol + ascites + ps + resdis +
log_ca125 + log_alp))
  # 補完後のデータセット（M組）の解析；Cox回帰

pool2 <- pool(ph2)
round(summary(pool2),3)
  # M組の解析結果の統合と結果の表示
```

Imputed Datasetの内訳

```
Multiply imputed data set
Call:
mice(data = tgce, m = 200, predictorMatrix = predmt1, seed = 34871)
Number of multiple imputations:  200
Missing cells per column:
    patno       age      figo     grade    histol   ascites        ps    resdis log_ca125   log_alp         d         t
        0         0        21       139         0        65       511        81       440       396         0         0
Imputation methods:
    patno       age      figo     grade    histol   ascites        ps    resdis log_ca125   log_alp         d         t
       ""        "" "polyreg" "polyreg"        ""  "logreg" "polyreg" "polyreg"     "pmm"     "pmm"        ""        ""
VisitSequence:
     figo     grade   ascites        ps    resdis log_ca125   log_alp
        3         4         6         7         8         9        10
PredictorMatrix:
          patno age figo grade histol ascites ps resdis log_ca125 log_alp d t
patno         0   0    0     0      0       0  0      0         0       0 0 0
age           0   0    0     0      0       0  0      0         0       0 0 0
figo          0   1    0     1      1       1  1      1         1       1 0 0
grade         0   1    1     0      1       1  1      1         1       1 0 0
histol        0   0    0     0      0       0  0      0         0       0 0 0
ascites       0   1    1     1      1       0  1      1         1       1 0 0
ps            0   1    1     1      1       1  0      1         1       1 0 0
resdis        0   1    1     1      1       1  1      0         1       1 0 0
log_ca125     0   1    1     1      1       1  1      1         0       1 0 0
log_alp       0   1    1     1      1       1  1      1         1       0 0 0
d             0   0    0     0      0       0  0      0         0       0 0 0
t             0   0    0     0      0       0  0      0         0       0 0 0
Random generator seed value:  34871
```

代入値の生成方法のオプション（R: mice）

Method	Description	Scale type	Default
`pmm`	Predictive mean matching	numeric	Y
`norm`	Bayesian linear regression	numeric	
`norm.nob`	Linear regression, non-Bayesian	numeric	
`mean`	Unconditional mean imputation	numeric	
`2L.norm`	Two-level linear model	numeric	
`logreg`	Logistic regression	factor, 2 levels	Y
`polyreg`	Multinomial logit model	factor, >2 levels	Y
`polr`	Ordered logit model	ordered, >2 levels	Y
`lda`	Linear discriminant analysis	factor	
`sample`	Random sample from the observed data	any	

Table 1: Built-in univariate imputation techniques. The techniques are coded as functions named `mice.impute.pmm()`, and so on.

van Buuren and Groothuis-Oudshoorn (2011)

Original Dataset

	patno	age	figo	grade	histol	ascites	ps	resdis	log_ca125	log_alp	d	t
1	1	45.67830	I	<NA>	Mucinous	present	0	<2 cm	4.852030	3.737670	0	7.1485284
2	2	70.10815	III	moderately differentiated	Serous papillary	absent	0	>5 cm	NA	4.330733	1	1.1690623
3	3	82.65572	III	poorly differentiated	Mesonephroid (clear cell)	present	<NA>	<2 cm	NA	NA	1	2.2532512
4	4	46.98426	III	poorly differentiated	Adenocarcinoma	absent	1	<2 cm	2.197225	4.234107	1	2.3983573
5	5	65.72485	IV	moderately differentiated	Serous papillary	present	0	>5 cm	NA	4.248495	1	1.4510609
6	6	38.33265	I	well differentiated	Serous papillary	present	<NA>	<2 cm	NA	NA	0	9.4565366
7	7	51.10746	III	poorly differentiated	Serous papillary	present	1	2-5 cm	5.347107	4.663439	1	1.0239562
8	8	63.72074	III	poorly differentiated	Serous papillary	absent	1	>5 cm	7.432484	5.187386	1	0.5585216
9	9	44.06845	I	poorly differentiated	Mucinous	absent	0	<2 cm	3.367296	4.317488	0	11.0171116
10	10	42.57358	III	poorly differentiated	Serous papillary	present	<NA>	>5 cm	NA	NA	1	2.3791923
11	11	60.49007	I	well differentiated	Endometrioid	present	0	<2 cm	NA	5.411646	1	13.2621492
12	12	56.50103	I	well differentiated	Endometrioid	present	1	<2 cm	2.944439	4.418840	1	6.9103354
13	13	67.51814	II	poorly differentiated	Mucinous	absent	0	<2 cm	3.637586	4.682131	1	4.5229295
14	14	57.42368	II	poorly differentiated	Serous papillary	absent	1	2-5 cm	7.989561	4.174388	1	1.1033539
15	15	45.28405	II	well differentiated	Serous papillary	present	0	<2 cm	4.697749	4.430817	1	1.5633128
16	16	71.59753	<NA>	<NA>	Adenocarcinoma	present	3to4	>5 cm	NA	4.574711	1	0.4982888
17	17	64.79398	I	moderately differentiated	Mucinous	present	<NA>	<2 cm	5.209486	NA	1	9.3032170
18	18	67.26900	III	moderately differentiated	Serous papillary	present	1	>5 cm	3.970292	4.189655	1	2.6967830
19	19	48.37235	III	poorly differentiated	Mucinous	absent	1	2-5 cm	5.945421	4.094345	1	1.7138946
20	20	57.04038	II	poorly differentiated	Serous papillary	present	0	<2 cm	NA	4.605170	1	5.3661875
21	21	59.67693	I	well differentiated	Endometrioid	absent	<NA>	<2 cm	NA	NA	0	10.4613279
22	22	43.38398	II	poorly differentiated	Serous papillary	absent	0	<2 cm	3.931826	4.488636	0	11.9041752
23	23	55.67967	I	moderately differentiated	Endometrioid	absent	0	<2 cm	2.484907	4.488636	1	0.8788501
24	24	76.51746	I	poorly differentiated	Serous papillary	absent	0	<2 cm	6.836259	NA	1	2.1574264
25	25	60.61328	I	well differentiated	Serous papillary	absent	<NA>	<2 cm	4.787492	NA	1	2.8966461

The 5th Imputed Dataset

	patno	age	figo	grade	histol	ascites	ps	resdis	log_ca125	log_alp	d	t
1	1	45.67830	I	moderately differentiated	Mucinous	present	0	<2 cm	4.852030	3.737670	0	7.14852841
2	2	70.10815	III	moderately differentiated	Serous papillary	absent	0	>5 cm	5.529429	4.330733	1	1.16906229
3	3	82.65572	III	poorly differentiated	Mesonephroid (clear cell)	present	1	<2 cm	4.158883	4.595120	1	2.25325120
4	4	46.98426	III	poorly differentiated	Adenocarcinoma	absent	1	<2 cm	2.197225	4.234107	1	2.39835729
5	5	65.72485	IV	moderately differentiated	Serous papillary	present	0	>5 cm	3.258096	4.248495	1	1.45106092
6	6	38.33265	I	well differentiated	Serous papillary	present	0	<2 cm	4.204693	4.189655	0	9.45653662
7	7	51.10746	III	poorly differentiated	Serous papillary	present	1	2-5 cm	5.347107	4.663439	1	1.02395619
8	8	63.72074	III	poorly differentiated	Serous papillary	absent	1	>5 cm	7.432484	5.187386	1	0.55852156
9	9	44.06845	I	poorly differentiated	Mucinous	absent	0	<2 cm	3.367296	4.317488	0	11.01711157
10	10	42.57358	III	poorly differentiated	Serous papillary	present	0	>5 cm	5.455321	4.488636	1	2.37919233
11	11	60.49007	I	well differentiated	Endometrioid	present	0	<2 cm	2.701361	5.411646	1	13.26214921
12	12	56.50103	I	well differentiated	Endometrioid	present	1	<2 cm	2.944439	4.418840	1	6.91033539
13	13	67.51814	II	poorly differentiated	Mucinous	absent	0	<2 cm	3.637586	4.682131	1	4.52292950
14	14	57.42368	II	poorly differentiated	Serous papillary	absent	1	2-5 cm	7.989561	4.174388	1	1.10335387
15	15	45.28405	II	well differentiated	Serous papillary	present	0	<2 cm	4.697749	4.430817	1	1.56331280
16	16	71.59753	IV	poorly differentiated	Adenocarcinoma	present	3to4	>5 cm	8.187577	4.574711	1	0.49828884
17	17	64.79398	I	moderately differentiated	Mucinous	present	0	<2 cm	5.209486	4.927254	1	9.30321697
18	18	67.26900	III	moderately differentiated	Serous papillary	present	1	>5 cm	3.970292	4.189655	1	2.69678302
19	19	48.37235	III	poorly differentiated	Mucinous	absent	1	2-5 cm	5.945421	4.094345	1	1.71389459
20	20	57.04038	II	poorly differentiated	Serous papillary	present	0	<2 cm	4.605170	4.605170	1	5.36618754
21	21	59.67693	I	well differentiated	Endometrioid	absent	0	<2 cm	4.127134	5.323010	0	10.46132786
22	22	43.38398	II	poorly differentiated	Serous papillary	absent	0	<2 cm	3.931826	4.488636	0	11.90417522
23	23	55.67967	I	moderately differentiated	Endometrioid	absent	0	<2 cm	2.484907	4.488636	1	0.87885010
24	24	76.51746	I	poorly differentiated	Serous papillary	absent	0	<2 cm	6.836259	5.780744	1	2.15742642
25	25	60.61328	I	well differentiated	Serous papillary	absent	0	<2 cm	4.787492	5.068904	1	2.89664613

Cox Regression for the 5th Imputed Dataset

```
Call:
coxph(formula = Surv(t, d) ~ age + figo + grade + histol + ascites + ps + resdis + log_ca125 + log_alp)

              coef exp(coef) se(coef)     z        p
age        0.02315   1.02342  0.00355  6.51 7.3e-11
figo2      0.68788   1.98948  0.16169  4.25 2.1e-05
figo3      1.19456   3.30211  0.13852  8.62 < 2e-16
figo4      1.23582   3.44119  0.16746  7.38 1.6e-13
grade2     0.24470   1.27724  0.15931  1.54  0.1245
grade3     0.27934   1.32226  0.15447  1.81  0.0705
histol2    0.23577   1.26588  0.19404  1.22  0.2244
histol3   -0.06239   0.93952  0.15968 -0.39  0.6960
histol4    0.35199   1.42190  0.13625  2.58  0.0098
histol5    0.82749   2.28757  0.18033  4.59 4.5e-06
histol6   -0.21044   0.81023  0.10158 -2.07  0.0383
histol7    0.22521   1.25259  0.24151  0.93  0.3511
ascites2   0.37842   1.45997  0.08382  4.51 6.3e-06
ps2        0.09476   1.09939  0.08883  1.07  0.2861
ps3        0.22873   1.25701  0.11694  1.96  0.0505
ps4        0.67439   1.96284  0.15084  4.47 7.8e-06
resdis2   -0.14265   0.86706  0.10427 -1.37  0.1713
resdis3   -0.65520   0.51934  0.10645 -6.15 7.5e-10
log_ca125  0.03206   1.03258  0.02736  1.17  0.2413
log_alp    0.37301   1.45210  0.07997  4.66 3.1e-06

Likelihood ratio test=744  on 20 df, p=0
n= 1189, number of events= 842
```

欠測値をImputeしたことによって、
すべてのデータを用いた解析を行うことができる

Cox Regression for the 25th Imputed Dataset

```
Call:
coxph(formula = Surv(t, d) ~ age + figo + grade + histol + ascites + ps + resdis + log_ca125 + log_alp)

              coef exp(coef) se(coef)     z       p
age        0.02261   1.02287  0.00354  6.38 1.7e-10
figo2      0.56968   1.76770  0.16207  3.51 0.00044
figo3      1.23025   3.42208  0.13846  8.89 < 2e-16
figo4      1.31465   3.72345  0.16522  7.96 1.8e-15
grade2     0.51168   1.66809  0.16442  3.11 0.00186
grade3     0.54567   1.72576  0.15930  3.43 0.00061
histol2    0.20187   1.22369  0.19102  1.06 0.29060
histol3   -0.00814   0.99189  0.15886 -0.05 0.95914
histol4    0.34285   1.40895  0.13779  2.49 0.01284
histol5    0.82787   2.28843  0.18256  4.53 5.8e-06
histol6   -0.16791   0.84543  0.10251 -1.64 0.10143
histol7    0.48894   1.63059  0.23860  2.05 0.04044
ascites2   0.30533   1.35708  0.08322  3.67 0.00024
ps2        0.04053   1.04136  0.08671  0.47 0.64021
ps3        0.14208   1.15267  0.12043  1.18 0.23808
ps4        0.68841   1.99054  0.14888  4.62 3.8e-06
resdis2   -0.16324   0.84939  0.10225 -1.60 0.11039
resdis3   -0.64448   0.52494  0.10337 -6.23 4.5e-10
log_ca125  0.01870   1.01888  0.02752  0.68 0.49687
log_alp    0.38169   1.46476  0.07808  4.89 1.0e-06

Likelihood ratio test=738  on 20 df, p=0
n= 1189, number of events= 842
```

Cox Regression for the 50th Imputed Dataset

```
Call:
coxph(formula = Surv(t, d) ~ age + figo + grade + histol + ascites + ps + resdis + log_ca125 + log_alp)

              coef exp(coef) se(coef)     z       p
age        0.02311   1.02338  0.00354  6.54 6.4e-11
figo2      0.64665   1.90913  0.16135  4.01 6.1e-05
figo3      1.20136   3.32463  0.13920  8.63 < 2e-16
figo4      1.23283   3.43093  0.16710  7.38 1.6e-13
grade2     0.32350   1.38195  0.16105  2.01 0.04457
grade3     0.45357   1.57392  0.15582  2.91 0.00361
histol2    0.11690   1.12401  0.19141  0.61 0.54137
histol3   -0.03584   0.96480  0.16119 -0.22 0.82405
histol4    0.28343   1.32768  0.13830  2.05 0.04042
histol5    0.76061   2.13958  0.18172  4.19 2.8e-05
histol6   -0.23698   0.78901  0.10187 -2.33 0.02000
histol7    0.35975   1.43297  0.24124  1.49 0.13589
ascites2   0.33703   1.40078  0.08308  4.06 5.0e-05
ps2        0.10642   1.11229  0.08629  1.23 0.21748
ps3        0.23393   1.26355  0.11923  1.96 0.04976
ps4        0.89309   2.44266  0.15248  5.86 4.7e-09
resdis2   -0.10031   0.90456  0.10355 -0.97 0.33268
resdis3   -0.67208   0.51064  0.10298 -6.53 6.7e-11
log_ca125  0.01057   1.01063  0.02702  0.39 0.69551
log_alp    0.26331   1.30123  0.07178  3.67 0.00024

Likelihood ratio test=739  on 20 df, p=0
n= 1189, number of events= 842
```

- 多重代入法（Multiple Imputation; MI）
- 欠測値に対して複数の代入値（M組）を生成
  - 代入値の生成方法はいろいろ（後ほど）
- M組の擬似的な完全データに対して、推定値と分散の推定値を求め、Rubinによる統合公式によって統合
  - $\hat{\beta}_h, \hat{V}(\hat{\beta}_h), h = 1,2,\ldots,M$

- 初期の教科書では、補完回数 M は3~5回で十分であるとされてきた（例えば、Rubin (1987)）
- もともと多重補完法が提案された1970~80年代では、十分な性能を持つ計算機がなく、Mを大きくしたもとでの計算は、現実問題として困難だという背景もあった
- Mを大きくとるほど、信頼区間・P値の近似精度は高くなる
  - 小さすぎると、近似精度が不十分である可能性が高い
- 最近の文献では、正確な推定・検定を行うために、100～1000回のオーダーにとることを勧めているものも多い

Carpenter and Kenward (2013), Royston and White (2011)

R example: MICE Output

```
> round(summary(pool2),3)
              est    se      t         df Pr(>|t|)  lo 95  hi 95 nmis   fmi lambda
age         0.024 0.004  6.436  77589.144    0.000  0.016  0.031    0 0.049  0.049
figo2       0.633 0.165  3.830  79236.415    0.000  0.309  0.956   NA 0.048  0.048
figo3       1.184 0.143  8.278  48394.035    0.000  0.904  1.465   NA 0.062  0.062
figo4       1.246 0.175  7.129  23140.643    0.000  0.904  1.589   NA 0.092  0.092
grade2      0.384 0.173  2.218  14749.187    0.027  0.045  0.723   NA 0.115  0.115
grade3      0.437 0.167  2.608  16488.140    0.009  0.109  0.765   NA 0.109  0.109
histol2     0.181 0.204  0.888  16983.350    0.375 -0.219  0.581   NA 0.107  0.107
histol3    -0.031 0.165 -0.188  59966.434    0.851 -0.354  0.292   NA 0.056  0.056
histol4     0.325 0.142  2.284  50458.739    0.022  0.046  0.603   NA 0.061  0.061
histol5     0.802 0.190  4.214  29427.975    0.000  0.429  1.174   NA 0.081  0.081
histol6    -0.206 0.104 -1.975 104717.958    0.048 -0.410 -0.002   NA 0.041  0.041
histol7     0.333 0.267  1.248   5984.292    0.212 -0.190  0.856   NA 0.182  0.182
ascites2    0.334 0.089  3.750  16858.463    0.000  0.159  0.508   NA 0.108  0.108
ps2         0.106 0.101  1.048   3096.892    0.295 -0.092  0.304   NA 0.253  0.253
ps3         0.207 0.144  1.443   2083.436    0.149 -0.074  0.489   NA 0.309  0.309
ps4         0.628 0.209  3.001    949.924    0.003  0.217  1.039   NA 0.458  0.457
resdis2    -0.116 0.112 -1.035  11215.134    0.301 -0.336  0.104   NA 0.132  0.132
resdis3    -0.650 0.112 -5.801   9575.774    0.000 -0.870 -0.431   NA 0.144  0.143
log_ca125   0.031 0.031  1.008   2997.766    0.314 -0.030  0.092  440 0.258  0.257
log_alp     0.331 0.096  3.428   1675.250    0.001  0.142  0.520  396 0.345  0.344
```

解析結果の比較①

	CCA (n=362, # deaths=248)				MICE (n=1189, # deaths=842)			
	HR	95%CI		P-value	HR	95%CI		P-value
Age (years)	1.02	1.00	1.03	0.011	1.02	1.02	1.03	< 0.001
FIGO stage								
I	1.00				1.00			
II	2.39	1.14	5.00	0.021	1.88	1.36	2.60	< 0.001
III	4.01	2.12	7.59	< 0.001	3.27	2.47	4.33	< 0.001
IV	5.83	2.81	12.09	< 0.001	3.48	2.47	4.90	< 0.001
Grade								
I	1.00				1.00			
II	1.28	0.65	2.52	0.471	1.47	1.05	2.06	0.027
III	1.13	0.59	2.13	0.717	1.55	1.11	2.15	0.009

†CCA: Complete-Case Analysis, MICE: Multiple Imputation by Chained Equation

解析結果の比較②

	CCA (n=362, # deaths=248)				MICE (n=1189, # deaths=842)			
	HR	95%CI		P-value	HR	95%CI		P-value
Histology								
Serous papillary	1.00				1.00			
Adenocarcinoma	1.54	0.37	6.36	0.554	1.20	0.80	1.79	0.375
Endometrioid	0.96	0.52	1.78	0.901	0.97	0.70	1.34	0.851
Clear cell	1.71	1.00	2.93	0.051	1.38	1.05	1.83	0.022
Mixed mesodermal	3.99	1.78	8.91	0.001	2.23	1.54	3.24	< 0.001
Mucinous	0.77	0.55	1.08	0.126	0.81	0.66	1.00	0.048
Undifferentiated	1.43	0.60	3.41	0.421	1.40	0.83	2.35	0.212
Ascites								
Absence	1.00				1.00			
Presence	1.35	0.98	1.86	0.069	1.40	1.17	1.66	< 0.001

†CCA: Complete-Case Analysis, MICE: Multiple Imputation by Chained Equation

解析結果の比較③

	CCA (n=362, # deaths=248)				MICE (n=1189, # deaths=842)			
	HR	95%CI		P-value	HR	95%CI		P-value
Performance status								
0	1.00				1.00			
1	1.16	0.86	1.57	0.325	1.11	0.91	1.35	0.295
2	0.97	0.62	1.50	0.885	1.23	0.93	1.63	0.149
3+4	2.30	0.96	5.48	0.061	1.87	1.24	2.83	0.003
Residual disease								
> 5 cm	1.00				1.00			
2-5 cm	0.97	0.67	1.41	0.880	0.89	0.71	1.11	0.301
< 2 cm	0.57	0.39	0.82	0.003	0.52	0.42	0.65	< 0.001
Log CA125	1.06	0.97	1.17	0.205	1.03	0.97	1.10	0.314
Log alkaline phos.	1.61	1.14	2.27	0.006	1.39	1.15	1.68	0.001

†CCA: Complete-Case Analysis, MICE: Multiple Imputation by Chained Equation

Some Remarks on MICE

- 近年の研究により、理論的妥当性が厳密に保証される条件はやや限定的であることが示されている
- ただ、多くの性能評価研究で、かなり広範な条件下で良好な成績が示されている（少なくとも、実践的には Favorable or Acceptableであることが多い）
- また、複数の変数にまたがって欠測が起こるような一般的な条件下で、汎用的なソフトウェアが充実している数少ない方法であり、現状の医学ジャーナルの多くのレビューでも、欠測データの統計解析に最も用いられているのは MICE であるとされている

Liu et al. (2013), Rezven et al. (2015), White et al. (2011)

まとめ

- 欧米によるガイドライン作成などを受けて、医薬品開発の臨床試験から、アカデミアで行われる臨床研究まで、不完全データの解析に求められる科学的妥当性の要求は高まっている
- 欠測データの解析方法は、近い将来、臨床研究・疫学研究におけるデータ解析のリテラシーのひとつとして求められていくことが予想される
- 詳細な解析方法についてご不安があるときには、統計家にご相談いただくことをお勧めします

Example Dataset & Program

- 本日のセミナーで用いた事例データと解析プログラムは、以下のURLからダウンロードできます
  - http://normanh.skr.jp/materials.html
- NMARのもとでの解析プログラムなども含めた、SASプログラム集は、『欠測データの統計科学』（高井啓二，星野崇宏，野間久史著；岩波書店）のホームページからダウンロードできます
  - http://www2.itc.kansai-u.ac.jp/~takai/Book/index.html

文献

- Carpenter, J., and Kenward, M. G. (2013). *Multiple Imputation and Its Application*. Chichester: Wiley.
- Clark, T. G., and Altman, D. G. (2003). Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *Journal of Clinical Epidemiology* **56**, 28-37.
- Clark, T. G., Stewart, M. E., Altman, D. G., Gabra, H., and Smyth, J. F. (2001). A prognostic model for ovarian cancer. *British Journal of Cancer* **85**, 944-952.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.
- Little, R. J., D'Agostino, R., Cohen, M. L., *et al.* (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* **367**, 1355-1360.
- Liu, J., Gelman, A., Hill, J., Su, Y.-S., Kropko, J. (2013). On the stationary distribution of iterative imputations. *Biometrika* **101**, 151-173.
- National Research Council. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, D.C.: National Academies Press.

- Rezvan, H. P., Lee, K. J., and Simpson, J. A. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology* **15**, 30.
- Royston, P., and White, I. R. (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software* **45**, Issue 4.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45**, 1-67.
- White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30**, 377-399.
- 日本製薬工業協会医薬品評価委員会データサイエンス部会．(2014)．臨床試験の欠測データの取り扱いに関する最近の展開と今後の課題について—NASレポート，EMAガイドライン，estimandと解析方法の概説．
http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/pdf/data_140704_1.pdf

- 日本製薬工業協会医薬品評価委員会データサイエンス部会．(2016)．欠測のある連続量経時データに対する統計手法について（Ver. 2.0）．
  http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/pdf/statistics01.pdf
- 高井啓二，星野崇宏，野間久史．(2016). 欠測データの統計科学：医学と社会科学への応用．岩波書店：東京.