厚生労働科学研究費補助金　難治性疾患政策研究事業
運動失調症の医療基盤に関する調査研究班　分担研究報告書

J-CATを活用した自然歴研究における欠測データの取り扱いについての考察

研究分担者　大西　浩文　（札幌医科大学医学部公衆衛生学講座）

研究要旨

J-CAT の登録数が順調に増え、登録データを活用した自然歴研究の検討が進んでいる。J-CAT 自然歴研究においては, SARA score をアウトカムとし、繰り返し測定データに基づく分析が想定されるが、観察研究の場合、十分な研究プロトコルの検討を行っても、患者背景要因も含めた未検査や脱落など予定しないデータ欠測が発生しうる。これまでの欠測値の取り扱いに関しては、欠測値を含む対象のデータを削除して分析する簡易な方法が適用されていたが、近年、欠測値を補完するモデルを用いて欠測値の補完を行い、欠測値が埋められた完全データを複数作成し、完全データに対して予定された解析を行い、得られた複数の結果を併合する手法である多重補完法 (multiple imputation; MI)が行われるようになった。MI 法は、平均値補完など単一の値を埋めて解析する手法と比較して、バラツキの過少評価を防ぐことができるため、欠測値を伴ったデータに対する手法として主要な手法の一つとなっている。今回は、本方法の今後の J-CAT 自然歴研究への応用が期待できる。

A. 研究目的

J-CAT の登録数が順調に増え、登録データを活用した自然歴研究が予定されている。観察研究の場合、十分な研究プロトコルの検討を行っても、患者背景要因も含めた未検査や脱落など予定しないデータ欠測が発生しうることから、本研究班での欠測データ取り扱いの可能性について検討する。

B. 研究方法

欠測データ発生のメカニズムとその対処方法についての方法論について利点と欠点などについてまとめ、J-CAT 自然歴研究における収集されるデータの特徴と用いられる分析方法について検討を行い、今後の応用の可能性を考察した。

(倫理面への配慮)

該当なし

C. 研究結果

観察研究の場合、十分な研究プロトコルの検討を行っても、患者背景要因も含めた未検査や脱落など予定しないデータ欠測が発生しうる。

これまでの欠測値の取り扱いに関しては、欠測値を含む対象のデータを削除して分析する簡易な方法が適用されていたが、近年、欠測値を補完する統計モデルを用いて欠測値の補完を行い、欠測値が埋められた完全データを複数作成し、完全データに対して予定された解析を行い、得られた複数の結果を併合する手法である多重補完法 (multiple imputation; MI)が行われるようになった [1)]。MI 法は、平均値補完など単一の値を埋めて解析する手法と比較して、バラツキの過少評価を防ぐことができるため、欠測値を伴ったデータに対する手法として主要な手法の一つとなっている。

データの欠測メカニズムとしては、① 完全にランダムな欠測 Missing Completely At Random (MCAR)、② ランダムな欠測 Missing At Random (MAR)、③ ランダムでない欠測 Not Missing At Random (NMAR, MNAR)の 3 種類が知られている。MCAR とは、欠測が完全にランダムに発生しており、モデリングの対象となる変数および関連する変数に依存しない場合であり、MAR は欠測す

るかどうかの確率が観測値には依存するが、欠測値には依存しない場合、NMAR は欠測するかどうかの確率が欠測値にのみ依存する場合である。

また、欠測データへの対処法としては、完全ケース分析（listwise deletion; LD 法）、利用可能なケースによる分析（pairwise deletion; PD 法）、完全情報最尤法（full information maximum likelihood method; FIML 法）、代入法の 4 種類が知られており、データ欠測のメカニズムによって、用いる対処方法も異なる。欠測データが MCAR の場合は、LD, PD は適用可能であるが、推定精度が落ちることが知られており、完全情報最尤法(FIML)や多重代入法だとこの問題を解決できると考えられる。MAR の場合は、LD, PD はバイアスが生じるので適用不可であり、FIML や多重代入を適用する必要がある。NMAR の場合は、欠測メカニズムの同定は一般に困難であり、限界はあるが MAR として分析を行うという選択となり、ベストな解決策がないのが現状である。

代入法には、単一代入法と多重代入法があるが、単一代入法としては平均値代入、回帰代入、確率的回帰代入、マッチングの 4 種類が知られている。平均値代入法とは、観測データ部分における平均値をそのまま代入値として採用する方法であり、結果にバイアスが含まれることが明らかとなっており推奨されない方法となっている。回帰代入は、欠測が起きていない変数で欠測値を予測する回帰モデルから得られた予測値を代入する方法であり、確率的回帰代入は、さらに変数が連続変数の場合には誤差項を考慮し、カテゴリーの場合は所属予測確率を使用する方法である。また、マッチングは、欠測している個体と欠測していない個体のマッチングを行い、欠測していない個体の値を代入する（傾向スコアマッチング）方法である。しかし、単一代入法ではデータ自体の影響を受けやすいことや、単一値を代入することによる精度の課題もある。そこで、単一代入法を複数回繰り返して疑似完全データを作成し、疑似完全データごとに計算した複数の推定値の統合を行う多重代入法が近年注目されている。

多重代入の具体的な手順としては、一般的に「代入ステージ」、「解析ステージ」、「統合ステージ」の 3 つのステージからなる。代入ステージにおいては、単一代入法で利用されている欠測値の代入法のいずれか、またはデータ拡大アルゴリズムなどを利用して D（≧2）個の疑似完全データを作成する。解析ステージでは、作成された D 個の疑似完全データから D 個の推定値を得る。最後の統合ステージにおいては、得られた D 個の推定値を統合することになる。

D. 考察

多重代入法の利点は、欠測データメカニズムの考慮や補助変数等の設定は代入実施者が行えばよく、欠測の取り扱いを解析者が行わなくてもよいという利便性と、代入モデルと解析モデルが必ずしも同じモデルである必要がないという柔軟性である。一方で欠点としては、代入モデルやそこでの推定法が誤っていれば、解析モデルが正しくても一致性のある推定量が得られない可能性があること、代入モデルとしてマッチングなどのノンパラメトリックな代入法を用いた場合の「統合された推定量」の数理的な性質が不明確であること、代入モデルでのみ説明力のある共変量を利用する場合などを除き、推定の効率性（推定量の標準誤差の小ささ）という点では Rubin 流の多重代入法よりも「直接尤度を最大化する最尤推定」の方が良いこと、Rubin のルールを利用した推定量の標準誤差の計算はデータセット D が小さいときにバイアスが存在することはソフトウェア開発者にはまだ浸透しておらず、多くのソフトウェアでそのまま出力されてしまうことなどが挙げられる。これらの利点、欠点を念頭においた上で多重代入法を適用する必要があると考えられる。

本研究班での J-CAT 自然歴研究では、SARA score を含む追跡データが脱落等によって欠損となる可能性があること、また特発性小脳失調症（IDCA）の調査において、主治医ごとに除外診断の検査実施に違いがあり、欠測が発生する可能性が考えられ、推定結果に無視できないバイアスを含むことがないように、適切な欠測の取り扱いが必要になる。したがって、今回検討を行った多重代入法は一つの解決策として検討の余地があると考えられる。また、代入するデータの推定精度を検討する上では、完全データからランダムに欠測を発生させるシミュレーションを行うなどの検討も必要になると考えられた。

E. 結論

J-CAT の自然歴研究や除外診断が中心となるIDCA の臨床背景および自然歴研究においては、欠測値が発生する可能性が十分考えられる。欠測値を適切に処理しない分析は推定結果に無視できないバイアスが生じることから、多重代入法は一つの解決策として検討の余地があり、J-CAT 自然歴研究への応用も期待できる。ただし、欠損のメカニズムによっては必ずしも期待される効果が得られない可能性も念頭に置く必要がある。

[参考文献]

1）Scheuren F. Multiple imputation: How it began and continues. The American Statistician 2005; 59: 315-319.

F. 健康危険情報

なし

G. 研究発表

なし

1.論文発表

なし

2.学会発表

なし