



2020年11月

データの管理について【第33回生物統計学】

1 概要

臨床試験では結果のデータに限らず試験スケジュールや対象者の情報など様々なデータが収集され、それぞれのデータシートに入力、保存されます。これらのデータシートを相互に結び付け、モニタリングや解析に使用できるクエリを作成するためにはデータベースを構築する必要があります。

2 データテーブル

データベースを構成するデータシートについて、行には対象者が、列には被験者の情報や測定データなどの変数となるデータが該当します。対象者には個人IDを割り振り、IDで対象者を認識することで個人情報の漏洩を防ぐことができます。一つのデータシートには一人の対象者が一行になるため、複数回測定するものや対象者の基本情報などでテーブルが複数になるものはテーブルごとに共通の個人IDを入力し、テーブル間の個人IDどうしを連結させることでデータベースを構築できます。このような複数のテーブルからなるデータベースを構築することを正規化といい、データの重複の可能性を減らすことやデータの検索、修正を容易にするなどのメリットがあります。

また、テーブルのデータディクショナリにはテーブルの列の変数が定義されたフィールドがあり、データディクショナリ自体がテーブルとなっています。データディクショナリはデータベース自体のデータを持っていることからメタデータと呼ばれ、フィールドでデータ型や入力可能な範囲などのルールを設定することで入力ミスの可能性を減らすことができます。

3 データの入力

臨床試験のデータ収集では症例報告書として紙媒体で収集され、コンピュータに入力されますがこの時、入力ミス等を確認するために二重入力等で入力データの確認を行う必要があります。データの収集が複数の施設で行われる場合、多くはそれぞれの施設からオンラインで直接データの入力を行います。データ用紙やUSB等の記憶媒体を送付し、1施設で入力する方法もあります。この場合、送付する情報は個人情報の削除やパスワードを用いる等によりデータの漏洩を防ぐ必要があります。

現在の臨床試験のデータ収集では多くがオンライン入力となっており、紙媒体からの入力に比べ入力ミスの減少、異常値の早期算出、携帯可能なワイヤレス装置での入力が可能等のメリットがあります。

4 データの抽出

複数のテーブルを結合しクエリを作成することで、入力されたデータから必要なデータや条件を満たすデータ、関数計算を行った値を表示させる事等ができるようになります。クエリには個人識別番号がないためテーブルの結合の時は共通のIDが用いられているか等結合後のデータに矛盾や誤りが無いことを確認す

る必要があり、データの矛盾や誤入力为了避免するためには、クエリの作成後にダミーデータ等を用いてデータの入力画面、データテーブル、クエリ画面全てに修正すべき点がないかを確認した後原票のデータを入力していくことで矛盾や誤りの可能性を減らします。

5 データの保護

研究対象者の個人情報保護は義務付けられており、そのデータベース以外では意味を持たない ID によって管理する必要があります。また、複数のテーブルでデータベースを構築している場合、個人識別情報は独立したテーブルで管理し許可された者だけがアクセスできるようにプロテクトされたサーバー上で管理する必要があります。

データベースは逸失してしまうことがないように定期的にバックアップを行い、ネットワークから切り離された環境で保存されなければならない、適切に保存されているか確認が必要になります。また、最終的なデータベースをアーカイブに保存する必要があります。

6 参考文献

・Stephen B Hulley et al. 医学的研究のデザイン. 木原雅子,木原正博訳. 第4版, 株式会社メディカル・サイエンス・インターナショナル, 2014, p.274-288.