医療データ科学実習グループワーク実施要項

1 概要

- 医療分野のデータの収集 · 解析 · まとめと報告フローを実践する．
  - 班分けは PandA にアップしてあるので自分の班を確認する．
- 最終回で，自分たちの解析で得られた結果に関するプレゼンテーションを行う．
- 解析対象のデータは各班が自分たちで選定して取得する．データリポジトリ（データを集めて公開している Web サイト）の候補を本資料の後半に提示するので，そのいずれかからデータを選ぶことをお勧めする．
  - 完全に自分たちでデータを選んできても良いが，著作権などの権利関係には十分留意すること．自信がなければ教員に相談する．
- 講義で扱っていない解析を行なっても良いが，設定した目的に合致したものであることは大前提とし，結果はできるだけわかりやすい図表などにまとめる．自信がなければ教員に相談する．

2 各グループのデータセット

1 班： Differentiated Thyroid Cancer Recurrence (UCI ML Rep) `https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence`

テーマ：

2 班： Diabetes Health Indicators Dataset (Kaggle) `https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset`

テーマ：

3 班： Regensburg Pediatric Appendicitis (UCI ML Rep) `https://archive.ics.uci.edu/dataset/938/regensburg+pediatric+appendicitis`

テーマ：

4 班： National Health and Nutrition Examination Survey `https://wwwn.cdc.gov/nchs/nhanes/Default.aspx`

テーマ： オーラルヘルス，コロナ前後 vs インフル前後での比較

5 班： WASH Benefits Bangladesh Cluster Randomized Trial (ClinEpiDB) `https://clinepidb.org/ce/app/workspace/analyses/DS_c56b76b581/new/variables/EUPATH_0035127/EUPATH_0044124`

6 班：

7 班： CRASH-2 (Vanderbilt Biostatistics Datasets) `https://hbiostat.org/data/repo/crash2`

テーマ：

8 班： SUPPORT study datasets (Vanderbilt Biostatistics Datasets) `https://hbiostat.org/data/repo/supportdesc`

3 グループワークのスケジュール（暫定）

6/17： 班分け，データの選定と取得，課題設定（何を明らかにすることを目指すのか），解析計画の作成（課題設定を実現するためにどのような解析をどのような手順で行うか）.

6/24： 解析計画に沿った解析（データの集計，グラフなどによる可視化，基本的な統計分析）.

7/1： 解析計画に沿った解析（データの集計，グラフなどによる可視化，基本的な統計分析）の続き．結果の出力と整理.

7/8： （前半にコーディングテストがあるので後半のみ）

前半（3 限）： コーディングテスト

後半（4 限）： 診断医学研究に関する補足（講義 · 演習形式）．解析計画に沿った解析（データの集計，グラフなどによる可視化，基本的な統計分析）の続き．結果の出力と整理.

7/15： 解析計画に沿った解析（データの集計，グラフなどによる可視化，基本的な統計分析），結果の出力と整理の続き．プレゼン資料の作成.

7/22： 最終プレゼン.

4 最終プレゼンについて

- プレゼンは各班で 20 分（準備 1 分，発表 15 分，質疑応答 5 分）.
  - 全 8 班の予定なので所要時間はスムーズに進行すれば $8 \times 21 = 168$ 分.
- 以下の内容を盛り込むこと
  - 分析対象としたデータの概要と背景（どのような目的で取得されたデータか，対象疾患は何か，どのような患者集団か，など）．データの取得経路と権利関係（著作権的に問題がないこと）も明示すること
  - データ分析の目的（何を明らかにすることを目的としてデータ分析を行なったのか）

  - データ分析の方法（目的を達成するためにどのような分析をどのような手順で行なったのか）
  - データ分析の結果（分析によってどのような結果が得られたのか）
  - 結果の考察（得られた結果は最初に設定した目的を明らかにしているか）
  - 各メンバーの貢献（それぞれ主に何を担当したか）
- 発表はメンバー全員で行い，各メンバーが最低 1 回は喋る時間を設ける．

5 データリポジトリの候補

- データは概要と背景が明示されているものを使うこと
  - リポジトリに data description が詳しく書いてあれば OK
  - リポジトリに data description が詳しく書いていない場合，description が取得できる元論文などがあるデータのみ使用可能とする（例えば下記の UCI Machine Learning Repository や Vanderbilt Biostatistics Datasets はこちらの可能性が高いので注意する）

5.1 Clinical Epidemiology Resources（おすすめ 1）

ClinEpiDB: `https://clinepidb.org/ce/app`

- 疫学研究のデータを統合したオープンアクセスのデータベース [1]
  - 複数の疾患を対象とした大規模研究データを収録
  - ペンシルベニア大学やジョージア大学などの複数の機関が共同で開発 · 運営
  - ブラウザ上でデータを視覚的に探索できる機能あり

5.2 UCI Machine Learning Repository（おすすめ 2）

`https://archive.ics.uci.edu/`

- カリフォルニア大学アーバイン校（University of California, Irvine）が運営する機械学習用のデータセットを公開しているリポジトリ
- データの信頼性が非常に高い．多くのデータセットは学術論文と関連付けられており，「このデータがどのような研究でどのように使われたか」を追跡することができる
- 比較的サイズが小さくきれいに整形（クリーニング）されているものが大半
- データセットは「Classification（判別）」「Regression（回帰）」「Clustering（クラスタリング）」といった機械学習のタスクごとに整理されているが，グループワークではこの分類は気にしなくても良い
- 注）data description に注意

5.3 Kaggle Datasets

https://www.kaggle.com/datasets?search=medical

- 世界最大のデータ分析コンペティションプラットフォーム Kaggle が提供しているデータセットリポジトリ
- CSV ファイルのような表形式データから画像，テキスト，音声データまで多岐にわたるデータが公開されており，データセットごとにライセンス（利用規約）が明記されているためどのような目的で利用できるかが分かりやすくなっている
  - Kaggle では研究や学習目的で利用できる医療分野のデータセットも多数公開されている（特に画像データが豊富だが，電子カルテ情報などの臨床データや健康調査データも利用可能）
  - Kaggle 公式サイト（https://www.kaggle.com/）でメールアドレスや google アカウントによる無料アカウント作成が必要

5.4 Vanderbilt Biostatistics Datasets

https://hbiostat.org/data/

- ヴァンダービルト大学生物統計学部が教育 · 研究目的で公開している多彩な統計解析用データセット集
- 臨床試験 · 観察研究（例 : GUSTO I，SUPPORT，PBC など）や一般統計例を含む実データを収載
- R パッケージ〈Hmisc〉経由やウェブサイトから RData ／ CSV ／ SAS ／ Stata 形式で取得可能で，各データには変数辞書と解析例を付属
- 回帰モデリングや生存解析などの講義 · 書籍（Frank Harrell 著 Regression Modeling Strategies 他）の再現教材として広く利用され，オープンライセンスで二次利用が容易
- 注）data description に注意

5.5 National Health and Nutrition Examination Survey

https://www.cdc.gov/nchs/nhanes/about/index.html

- 米国疾病管理予防センター（CDC）が実施する国民健康 · 栄養調査の公開データベース（実際のデータが置いてあるページ : https://wwwn.cdc.gov/nchs/nhanes/Default.aspx）
- 1960 年代に開始し，1999 年以降は 2 年ごとに約 5,000 人を抽出して継続的に実施
- 身体測定 · 臨床検査 · 栄養インタビュー · 行動調査などを組み合わせ，健康状態 · 食事摂取 · 環境曝露を多面的に把握
- データファイル（SAS XPT 形式など）と技術文書を無償でダウンロード可能．ウェイト変数や結合解析方法の解説，オンラインの可視化ダッシュボードや API も提供されている
- 注）巨大なデータベースなのでデータの切り出し作業が大変

6 データセットの著作権の例

CC BY 4.0（Creative Commons Attribution 4.0 International）ライセンス

- 商用 · 非商用を問わず自由に利用可能
- 改変 · 加工 · 再配布も可能
- クレジットを明記することが必須
  - 作者名（提供者名），ライセンス名，リンク先などを明示する必要ある
    例（論文や Web サイトで使用する場合）:
    This dataset is based on data from [Dataset Name] by [Author/Organization], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).
  - 改変した場合はその旨を明記
    例：Modified from the original dataset. Original: [Dataset Name] by [Author], licensed under CC BY 4.0.

References

[1] E Ruhamyankaka, BP Brunk, G Dorsey, OS Harb, DA Helb, J Judkins, JC Kissinger, B Lindsay, DS Roos, EJ San, CJ Stoeckert, J Zheng, and SS Tomko. Clinepidb: an open-access clinical epidemiology database resource encouraging online exploration of complex studies. *Gates Open Research*, 3(1661), 2020.