CFAR Chalk Talk

Big Data Analytics in HIV/AIDS Clinical Research

Cliburn Chan and Shein-Chung Chow
CFAR Biostatistics and Computational Biology Core
MSRB I, Duke University School of Medicine

June 29, 2017
Outline of Part I

• What is big data analytics?
  • Definition & purpose
  • US NIH BD2K initiative
  • Promise and potential

• HIV/AIDS database
  • Data quality and validity
  • FDA Part 11 compliance for electronic records

• Challenging statistical issues
  • Statistical methods
  • An example of case-control studies

• Future perspectives
What is big data analytics?

• Big data analytics is referred to as the analysis of **large data sets** from various data structured, semi-structured, or unstructured **sources** from healthcare related biomedical research
  – Registry studies
  – Randomized or non-randomized studies
  – Published or unpublished studies
  – Healthcare data (hospital, insurance, etc)
  – Healthcare networks
  etc
Purpose

• The purpose of big data analytics is to detect any possible hidden
  – signals
  – patterns and/or trends
  – risk/benefit
  – predictive medical model
of safety and efficacy of certain test treatments under study
• The information can be used for future planning of clinical trials
US NIH BD2K initiative

- BD2K = Big Data to Knowledge
- Biomedical research is rapidly becoming data-intensive as investigators are generating and using increasingly large, complex, multi-dimensional and diverse data sets.
- NIH RFP calls for
  - Statistical methods and software tools for data analysis
    - Data compression and reduction
    - Data provenance and wrangling
Promise and potential

• Big data analytics provides the following opportunities
  – Uncover hidden medical information
  – Determining possible associations or correlations between potential risk factors and clinical outcomes
  – Predictive model building, validation and generalization
  – Data mining for biomarker development
  – Critical information for future clinical trials planning
HIV/AIDS database
Data quality and validity

• Accepting data
  – Confidentiality
  – Agreement for data sharing
  – Development of standard forms for data capture

• Data management
  – Data transfer
  – Data review/query
  – Data verification/validation
  – Database lock
  etc
Figure 1: Typical example of data management process.
US FDA Part 11 compliance for electronic records

• Big data center usually contains data (electronic records) from various sources of structured, semi-structured or unstructured sources


• These regulations and requirements are known as Part 11 compliance
US FDA Part 11 compliance for electronic records

• Part 11 compliance plan
  – Gap analysis
  – User requirements specification
  – Validation master plan
  – Tactical implementation plan
  – Part 11 checklist
  – Final report
Challenging statistical issues

• Representativeness of big data
  – Selection bias
  – Heterogeneity
  – Reproducibility and generalizability
  – Missing or incomplete data

• Data quality, integrity, and validity
  – Criteria for accepting data
  – Confidentiality issue
  – Data management
  – Data transfer
  – Data sharing
Challenging statistical issues

• Statistical methodology
  – Case-control study
  – Meta-analysis
  – Data mining (genomics studies)

• Software development
  – Propensity score
  – Predictive model building, validation, and generalization
  – Variable screening
Representativeness of big data

• For a given disease under study, it is a concern whether big data is representative of the target patient population due to the fact that big data contains data from individual studies with
  – Similar but different study protocols with similar but different inclusion/exclusion criteria
  – Similar but different target patient populations
  – Similar but different study objectives/hypotheses/endpoints
  – Similar but different trial procedures
  – Similar but different statistical procedures
  etc
Selection bias

\[ \mu = \text{true mean of target patient population} \]
\[ \mu_B = \text{true mean of the big data} \]
\[ \mu_P = \text{true mean of data sets with positive results} \]
\[ \mu_N = \text{true mean of data sets with negative results} \]
\[ r = \text{true proportion of data with positive results} \]

\[ E(\hat{\mu}_B) = \mu_B = r \mu_P + (1 - r) \mu_N \]
\[ \text{Bias}(\hat{\mu}_B) = E(\hat{\mu}_B) - \mu = \mu_B - \mu = \varepsilon \]
Selection bias

\[ \text{Bias}(\hat{\mu}_B) = E(\hat{\mu}_B) - \mu = \mu_B - \mu = \varepsilon \]

- The above leads to \( \varepsilon = (1 - r)(\mu_P - \mu_N) \)
  - This could be substantial if there is a major difference between \( \mu_P \) and \( \mu_N \).
  - In practice, \( r \) is usually unknown
- If big data only contains data sets with positive results, then \( \mu_B = \mu_P \)
Heterogeneity

• Similarities/differences within and across individual studies with
  – Different means
  – Different variances
  – Different sample sizes
• Possible treatment-by-study interaction
  – Poolability for final analysis
• Possible confounding effects
  – Baseline demographics
  – Patient characteristics
Confounding factors

• **Baseline demographics**
  – Age
  – Gender
  – Weight/height
  – Race
  etc

• **Patient characteristics**
  – Disease severity
  – Medical history
  – Concomitant medication
  etc
Interactions

- Effect due to interactions between
  - Treatment
  - Center
  - Covariates (demographics, patient characteristics)
- Data should not be pooled for analysis if a significant qualitative interaction is observed
- Data may be pooled for final analysis if a significant quantitative interaction is observed
- In practice, it is suggested that tests for possible interactions should be performed
Missing data

- Missing data or incomplete data are commonly encountered in biomedical research
  - Dropouts
  - Lost to follow-up
  - Withdraw of informed consent
  - Withdraw by investigators
    etc
- How to handle missing data or incomplete data?
  - Completer analysis or missing data imputation?
  - Control of overall type I error rate
  - Achieving desired power
Reproducibility/genealizability

• Reproducibility
  – From one big data center (e.g., UNC Chapel Hill Medical Center) to another big data center (e.g., Duke University Medical Center)

• Generalizability
  – From one target patient population (e.g., adults) to another similar but different target patient population (e.g., children or elderly)
Statistical methods

• Case-control studies
  – Propensity score
  – Multivariate (logistic) regression analysis
  – Model building, validation, and generalization

• Meta analysis
  – Test for treatment-by-study interaction
  – Similarities and dis-similarities

• Data mining
  – Variable screening
  – Qualification
  – Validation
Case-control studies
Steps for model building

- Propensity Score:
  - Probability of a subject being assigned to a particular treatment
  - Construct matched pairs

- Model Building:
  - Association of patient characteristics and outcome
  - Prepare variables for modeling

- Model Selection:
  - Include multiple variables in the model
  - Select the best subset of variables

- Model Validation:
  - Use internal validation to assess the model selected
Observational studies – What is the problem?
Observational studies – What is the problem?

• Are we comparing apples to oranges?

\[ T = 0 \] (receive treatment A)  
Events: 20 out of 1000

\[ T = 1 \] (receive treatment B)  
Events: 5 out of 1000

**Question:** Is treatment B *better* than treatment A?

**Answer:** not necessarily true.

*Potential Confounding factors:* age, gender, overall health condition, and *etc.*

For Example:  
Treatment A: average age is 45 years old.  
Treatment B: average age is 25 years old.
Propensity score for matching control

- **Definition**: the *conditional probability* of a unit (i.e., patient) being assigned to a particular treatment given a set of observed covariates (i.e., baseline characteristics: age and gender).
  - $p(X) = \Pr[T = 1 | X]$

- Rosenbaum and Rubin (1983): treatment assignment to be strongly ignorable if:
  - (a) $(R_1, R_0) \perp T | X$
  - (b) $0 < p(X) < 1$
Propensity Score

• **Advantages**
  – Useful when adjusting for a large number of risk factors
  – Balances treatment and control groups.

• **Disadvantages**
  – Requires large samples
  – Only accounts for observed (and observable) covariates, so hidden bias may remain
Model building – univariate analysis

- Review the relationship of the outcome with each predictor

\[ \frac{\pi}{1-\pi} = \exp(\alpha + \beta x) = e^\alpha e^{\beta x} \]

- \( \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x \)

- Test of significance (Wald Test):
  - \( H_0: \beta = 0 \)
  - \( Z = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim N(0,1) \)
  - \( H_0 \) would be rejected when \( Z > Z_{1-\alpha/2} \) for a two-sided test at \( \alpha \) level
Model building - collinearity analysis

- Concern of variables could be correlated with one another
- Continuous variables: Correlation matrix
- Binomial/multinomial variables: Chi-square test
- Variance Inflation Factor (VIF):
  - \[ VIF_j = \frac{1}{1-R_j^2}, \quad j = 1,2,3, \ldots, k \] where \( R_j^2 \) is the squared multiple correlation based on regression \( X_j \) on the remaining \( k-1 \) predictors
  - Typical rule of thumb: VIF >10 indicates the existence of collinearity issue
Model building - multivariate analysis

- $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

- Global null hypothesis testing:
  - $H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$
  - $G = -2\ln \left(\frac{\text{Likelihood of the null model}}{\text{Likelihood of the full model}}\right) \sim \chi^2_{1-\alpha/2(k)}$

- Model selection
  - Wald statistic of significant variables
  - Forward, backward, and stepwise procedure
  - AIC to select the most appropriate model: $\text{AIC} = -2\ln L + 2p$
Model diagnostic

- Deviance goodness-of-fit test:
  - $H_0$: the model is correct for the dataset
  - $D = 2[l(\hat{\pi}, y) - l(\tilde{\pi}, y)]$
  - $D \sim \chi^2(n-p)$ where $p$ is the number of predictors in the model of interest
  - $H_0$ would be rejected when $D > \chi^2_{1-\alpha/2}(n-p)$
Model validation

• Data set for model validation: randomly split the given dataset
  – Typical approach: 90% as training data and 10% as test data
• Training set: build and refine the predictive model
• Test set: validate the preliminary final model
• Criterion:
  – Akaike’s information criteria (AIC): $\text{AIC} = -2\ln L + 2p$
Model generalizability

Let
\[(\mu_0, \sigma_0) = \text{target population of the big data center}\]
\[(\mu_1, \sigma_1) = \text{another target population of the big data center}\]

Since patient population in other big data center is similar but different, it is reasonable to assume
\[
\mu_1 = \mu_0 + \varepsilon
\]
\[
\sigma_1 = C \sigma_0
\]

Where \(\varepsilon\) is a shift in population mean and \(C\) is a inflation factor of population standard deviation
Model generalizability

There is a relationship between the effect size adjusted for standard deviation between the original big data center and the other big data center

\[ E_1 = \frac{\mu_1}{\sigma_1} = \frac{\mu_0 + \varepsilon}{C \sigma_0} = |\Delta| E_0 \]

where

\[ \Delta = \left[ 1 + \frac{\varepsilon}{\mu_0} \right] / C \]

in which, $\Delta$ is so-called sensitivity index
Controversial issue #1

• The finding of the big data analytics is **inconsistent** with that of from a relatively small scale of adequate well-controlled randomized clinical trial which was conducted under the similar target patient population

• Which result (conclusion) is reliable?
  – If the result from the big data analytics is considered more reliable, this may suggest that there are no need to conduct randomize clinical trials in the future.
  – If the result from the small scale randomized clinical trial is more reliable, this may suggest the useless of big data analytics
Comments

• The representativeness of the big data may be questionable which may be due to
  – Accepting poor data sets in the big data
  – Selection bias
  – Heterogeneity or dissimilarities across individual data sets in the big data

• It is then suggested that the finding of the big data analytics be appropriately adjusted
  – Further research is needed.
Controversial issue #2

• The finding of big data analytics using a specific statistical method is inconsistent with the one obtained using similar but slightly different method
  – For example, logistic regression analysis with forward stepwise approach and logistic regression analysis with backward stepwise approach
  – It has been a concern that the result from big data analytics is not reproducible
Comments

• An example
  – In practice, it is very possible that the logistic regression analysis with forward stepwise approach identifies $x_1, x_2,$ and $x_3$, while the logistic regression analysis with backward stepwise approach identifies $x_1, x_2,$ and $x_4$
  – Thus, researchers might claim that the result from the big data analytics is not reproducible
  – In this case, $x_3$ may be correlated with $x_4$. Thus, it is suggested that a composite index, say $y_3$ be developed. In other words, $y_3 = f(x_3, x_4)$. 
Controversial issue #3

• For big data analytics, in practice, it is likely that the findings at different time periods are different due to the fact that
  – the availability of the advanced technology
  – genetic changes in patient population
  – healthcare over time
• There is a possible time trend in data sets in the big data
Comments

- It is suggested that the following factors (expected to change over time)
  - the availability of the advanced technology
  - genetic changes in patient population
  - healthcare
be taken into consideration in the statistical model for a more accurate and reliable assessment of treatment effect (or clinically meaningful difference) under investigation.
Future perspectives

• Big data analytics is helpful in uncovering some hidden medical information but there are some practical issues and limitations

• Quality, validity, and integrity of the big data is the key to the success of the big data analytics

• Big data analytics in biomedical research should not be misused or abused