ALEH Clinical Research Workshop: How to Use and Manage Databases

David Goldberg, MD, MSCE
Assistant Professor of Medicine
Assistant Professor of Epidemiology
Senior Fellow, Leonard Davis Institute of Health Economics
Outline for today

• Brief introduction to using large databases
• Discuss pros and cons of large databases
• Overview of how to choose the right database and potential databases for research
• How to analyze and interpret large databases
• Pitfalls and helpful tools
Basics of using large databases

• What is a database?
  – Collection of data that is organized so that its contents can easily be accessed, managed, and updated

• Database sizes
  – Small: $<10^5$ records, $<10$GB data
  – Medium: $10^5$-$10^7$ records, 10-40GB data
  – Large: $>10^7$ records, $>40$GB data

• Records ≠ patients
  – Patients may have multiple entries
  – Multiple updates (i.e., MELD updates)
Major logistical issues to consider with database research

• Access to statistical software and/or statistical support
  – SPSS is point-and-click
  – Having software ≠ understanding statistics

• Can my computer handle the data (10MB->1GB)

• Do I have/need funding for the data
  • UNOS and SRTR transplant databases: $250-2,500

• Am I able to clean data
  – Why was data collected?
  – Missing data
  – Repeat entries
Why use a large database

• When a large sample size needed (rare exposure or outcome)
• Compare outcomes/performance across some measure
  – Variation in transplant center post-OLT outcomes
  – Organ donation rates across donor service areas
• Weigh benefits and tradeoffs of large database vs single-center data
  – Loss of granularity (can’t review medical records for 100,000 people)
  – Lack of control for data entry (previously coded or administrative data)
• Potential studies evaluating outcomes of cirrhotics in ICU
  – Single center: Outcomes, reason for admissions, risk factors (MELD, APACHE, SOFA) for adverse outcomes
  – Large database (PHC-4): All cirrhotics in ICUs in PA
    • Evaluate outcomes and reasons for admissions
    • Compare outcomes across hospitals (academic vs community)
    • Don’t have lab data data (MELD, SOFA)
  – What is the main question/message
Pros of using large databases

• Large sample size
  – Never underpowered (10:1 ratio outcomes:covariate)
  – Easy to get statistical significance

• Generalizability and external validity
  – Usually capture robust population
  – Single-center study not generalize to broader population

• Can be geographically and demographically diverse
  – Single-center vs national data

• Compare data across areas (geography, centers)
Cons of using large databases

- Large sample size: Easy to get statistical significance
- Data may be unreliable: Depends on who entered data
- ICD-9 codes (billing) in insurance
- Karnofsky score in UNOS
- Limited lab data
- Costs: Unruly and challenging to deal with

Hazard ratio: 0.89, 95% CI: 0.81-0.98; p<0.001
Choosing the right database

• Depends on:
  – Research question
  – Population of interest
  – Time
  – Budget

• Large database may not be the right answer

• Question and database:
  – Question: What are post-OLT outcomes of patients with PSC
    • Database: UNOS/SRTR
  – Question: What is the success rate of HCC downstaging protocols in the United States
    • Database: UNOS/SRTR (granularity), SEER (no Milan/UCSF), single/multi-center
  – Question: Are there differences in waitlisting for transplant across the United States
    • Database: ???—what is denominator
Using large database vs single-center data

- Depends on research question
- Are all the data available in both datasets
- Examples:
  - Does pre-transplant chronic kidney disease predict post-transplant survival
    - Single-center data better
      - Need to define CKD (i.e., renal ultrasound, proteinuria, trends)
  - Are increasing age and BMI associated with higher risks of early graft failure
    - Large database->more robust numbers
Transplant databases: UNOS and SRTR

- What is the OPTN?
  - Maintains the national registry for organ matching based on NOTA

- What is UNOS?
  - Private non-profit organization that has OPTN contract
  - Responsible for organ matching and collection of data

- What is the SRTR?
  - Organization responsible for analyzing transplant data, creating program-specific reports for center performance and public dissemination
  - Carries out analyses requested by OPTN committees

- SRTR and UNOS
  - Similar data
  - SRTR data “cleaned”
  - Different costs
  - Different request process
Interpreting results of large databases

• It’s not all about the p-value
  – P-value measures likelihood of finding something by chance
  – Largely influence by sample sizes
• Is it clinically meaningful
  – Don’t just look at HR/OR
  – Look at actual numbers and predicted outcomes
• Does the result make biological sense or just statistical anomaly (1/20 happen by chance)
Interpreting results: Hypothetical example

- Is the difference in outcomes really that large
- Research question: Is the 1-year post-OLT survival different for LT recipients with PSC vs PBC vs AIH
- Outcome: 1-year post-OLT survival (binary)

```
.xi: logistic died_within_1 i.psc_pbc_aih age dri ldlt final_meld_peld_lab_score i.psc_pbc_aih _Ipsc_pbc_a_0-2 (naturally coded; _Ipsc_pbc_a_0 omitted)

Logistic regression
Number of obs = 6,063
LR chi2(6) = 82.56
Prob > chi2 = 0.0000
Log likelihood = -1800.7012 Pseudo R2 = 0.0224

| died_within_1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval] |
|---------------|-------------|-----------------|---|-----|-----------------------------|
|                |             |                 |   |     |                            |
| AIH           | 1.383664    | .1488621        | 3.02| 0.003| 1.120608 - 1.70847          |
| PBC           | .9499216    | .1086927        | -0.45| 0.653| .7590859 - 1.188734         |
| age           | 1.023072    | .003816         | 6.12| 0.000| 1.015620 - 1.030579         |
| dri           | 1.206117    | .1334843        | 1.69| 0.090| .9709222 - 1.498284         |
| ldlt          | 1.398781    | .2576449        | 1.82| 0.068| .9749108 - 2.006941         |
| final_meld_peld_lab_score | 1.026813    | .0047798        | 5.68| 0.000| 1.017487 - 1.036224         |
| _cons         | .0111116    | .0032352        | -15.45| 0.000| .0062797 - .0196612         |

margins if psc_pbc_aih==0
Predictive margins Number of obs = 2,705
Model VCE : OIM Expression : Pr(died_within_1), predict()

| Margin   Std. Err.      z    P>|z|     [95% Conf. Interval] |
|----------|-----------------|---|-----|-----------------------------|
|          |                 |   |     |                            |
| _cons    | .0783734        | .0051373 | 15.26| 0.000| .0683045 - .0884423         |

margins if psc_pbc_aih==1
Predictive margins Number of obs = 1,662
Model VCE : OIM Expression : Pr(died_within_1), predict()

| Margin   Std. Err.      z    P>|z|     [95% Conf. Interval] |
|----------|-----------------|---|-----|-----------------------------|
|          |                 |   |     |                            |
| _cons    | .11071          | .0076355 | 14.50| 0.000| .0957448 - .1256752         |

margins if psc_pbc_aih==2
Predictive margins Number of obs = 1,696
Model VCE : OIM Expression : Pr(died_within_1), predict()

| Margin   Std. Err.      z    P>|z|     [95% Conf. Interval] |
|----------|-----------------|---|-----|-----------------------------|
|          |                 |   |     |                            |
| _cons    | .0902123        | .006928 | 13.02| 0.000| .0766336 - .103791          |
```
Analyzing UNOS data: You get your STAR file—now what?

• Step 1: Look at the data
• Step 2: Look at the data dictionary
  – What are the different variables?
  – How are the variables coded
• Code the data in an analyzable way
  – Gender (M/F) into 0s and 1s
• Evaluate for repeat and missing data
Knowing the lingo of UNOS

- **wl_id_code vs pt_code**
  - **pt_code**
    - One code per patient
    - Tracks through all waitlist entries
  - **wl_id_code**
    - One code per waitlist entry
    - Can have multiple codes (e.g., dual listing, re-transplant)

- **TCR vs TRR**
  - **TCR=transplant candidate registration**
    - Data at time of waitlisting
  - **TRR=transplant recipient registration**
    - Data at time of transplant
Conclusions and take-home points

• Large databases can be wealth of information
• Large sample sizes allow for important questions to be answered
• Need to be aware of limitations of databases
  – Validity of codes
  – Missing data
  – Lack of labs
  – Know what data initially created for
• Don’t get scooped—anyone can access UNOS data