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

CCEB

Penn Transplant Institute
Penn Medicine

Perelman School of Medicine
University of Pennsylvania

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, <10GB data
  - Medium: 10^5-10^7 records, 10-40GB data
  - Large: $>10^7$ records, >40GB 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

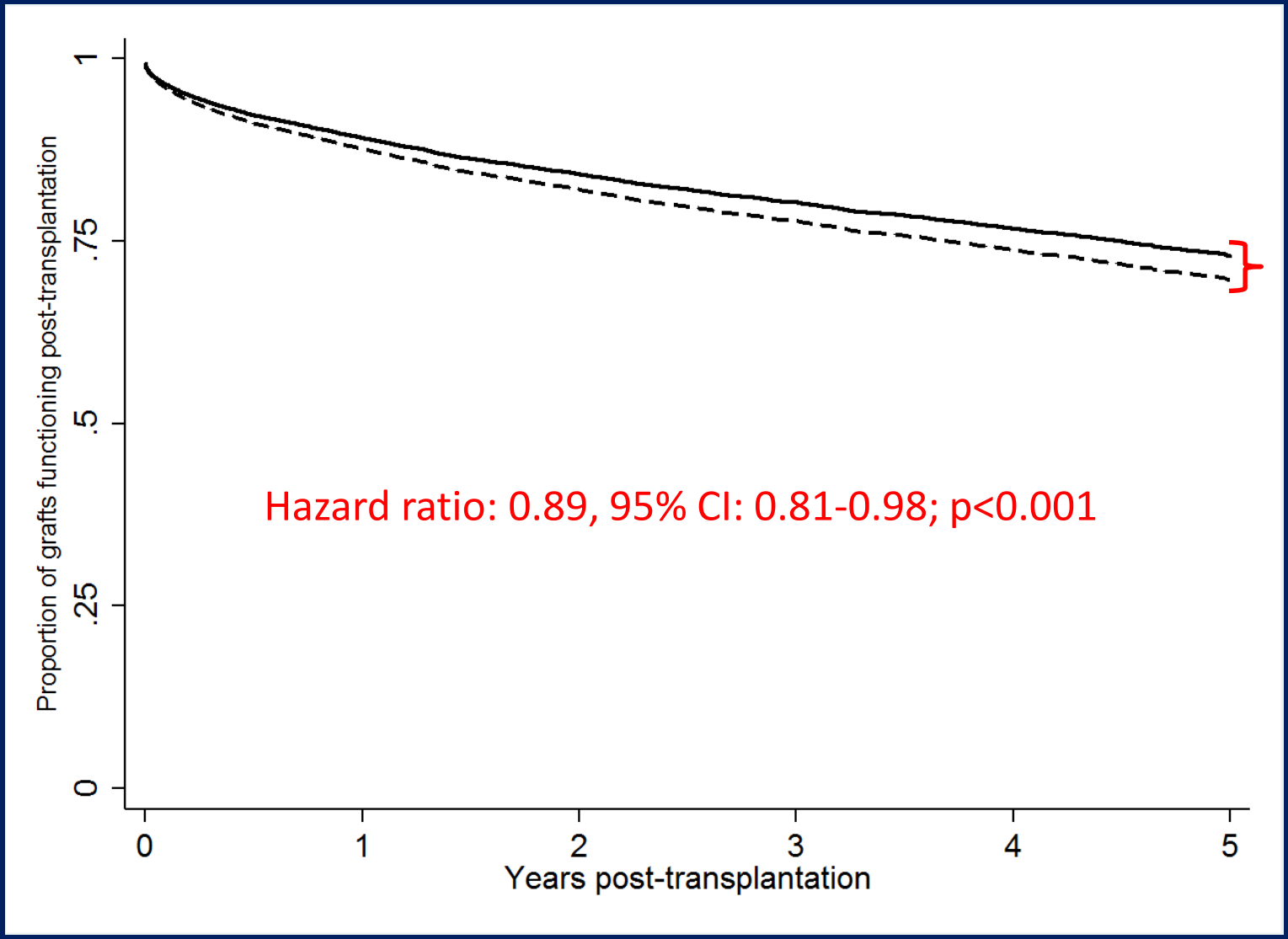
Why use a large database

- 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



4% absolute difference

Choosing the right database

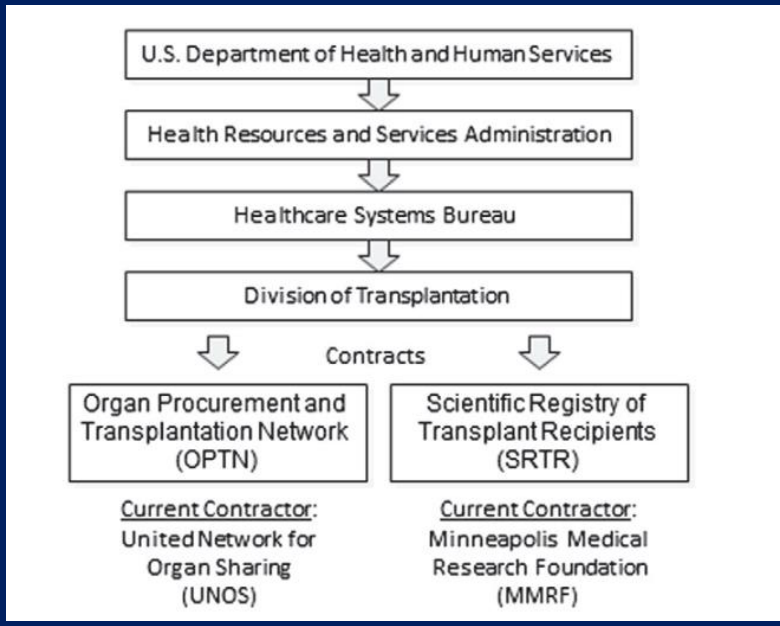
- Depends on:
  - Research question
  - Population of interest
  - Time
  - Budget
- Large database may not be the right answer
- Question and database:
  - Question: What are post-OLT outcomes of patients with PSC
    - Database: UNOS/SRTR
  - Question: What is the success rate of HCC downstaging protocols 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      1.01562           1
          dri   |   1.206117   .1334843     1.69   0.090     .9709222           1
         ldlt   |   1.398781   .2576449     1.82   0.068     .9749108           2
final_meld_peld_lab_score |   1.026813   .0047798     5.68   0.000     1.017487           1
        _cons   |   .0111116   .0032352   -15.45   0.000     .0062797           .
------------------------------------------------------------------------------
```

```
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?

-

	wl_id_code	pt_code	rem_cd	age	init_age	region	gender
1	571463	365907	15	28	25	11	F
2	1161229	1035475	4	60	60	11	M

	VARIABLE NAME	DESCRIPTION	FORM	VAR START DATE	VAR END DATE	FORM SECTION
3	ABO	RECIPIENT BLOOD GROUP @ REGISTRATION	TCR	01-Oct-87		CLINICAL INFORMATION
4	ABO_DON	DONOR BLOOD TYPE	DDR/LDR	01-Oct-87		DONOR INFORMATION
5	ABO_MAT	DONOR-RECIPIENT ABO MATCH LEVEL	CALCULATED			
6	ACADEMIC_LEVEL_TCR	ACADEMIC ACTIVITY LEVEL AT LISTING	TCR	30-Jun-04		CANDIDATE INFORMATION
7	ACADEMIC_LEVEL_TRR	ACADEMIC ACTIVITY LEVEL AT TRANSPLANT	TRR	30-Jun-04		PATIENT STATUS
8	ACADEMIC_PRG_TCR	ACADEMIC PROGRESS AT LISTING	TCR	30-Jun-04		CANDIDATE INFORMATION
9	ACADEMIC_PRG_TRR	ACADEMIC PROGRESS AT TRANSPLANT	TRR	30-Jun-04		PATIENT STATUS
10	ACUTE_REJ_EPI	ACUTE REJECTION EPISODE BETWEEN TRANSPLANT AND DISCHARGE?	TRR	30-Jun-04		POST TRANSPLANT CLINICAL INFORMATION
11	ACYCLOVIR	Biological or Anti-Viral Treatment - Acyclovir	TRR	30-Jun-04	01-Jan-07	Treatment
12	ADMISSION_DATE	RECIPIENT DATE OF ADMISSION TO TX CENTER	TRR	25-Oct-99		PATIENT STATUS
13	ADMIT_DATE_DON	DONOR ADMIT DATE	DDR	26-Apr-06		DONOR INFORMATION
14	AGE	RECIPIENT AGE (YRS)	TRR-CALCULATED	01-Oct-87		RECIPIENT INFORMATION
15	AGE_DON	DONOR AGE (YRS)	DDR/LDR-CALCULATED	01-Oct-87		DONOR INFORMATION
16	AGE_GROUP	RECIPIENT AGE GROUP A=ADULT P=PEDS	CALCULATED			
17	ALBUMIN_DIS	RECIPIENT SERUM ALBUMIN @ DISCHARGE	TRR	01-Oct-87	01-Jan-07	POST TRANSPLANT CLINICAL INFORMATION
18	ALBUMIN_TX	RECIPIENT SERUM ALBUMIN @ TRANSPLANT	WAITING LIST DATA	01-Oct-87		PRETRANSPLANT CLINICAL INFORMATION - SERUM LAB DATA
19	AMIS	A Locus MISMATCH LEVEL	CALCULATED			
20	ANGINA	RECIPIENT ANGINA/CAD @ REGISTRATION	TCR	30-Jun-04	01-Jan-07	CLINICAL INFORMATION
21	ANGINA_OLD	RECIPIENT ANGINA/CAD @ REGISTRATION	TCR	01-Apr-94	30-Jun-04	CLINICAL INFORMATION
22	ANTICONV_DON	DECEASED DONOR-ANTICONVULSANTS W/IN 24 HRS PRE-CROSS CLAMP	DDR	01-Apr-94		CLINICAL INFORMATION
23	ANTIHYPE_DON	DECEASED DONOR-ANTIHYPERTENSIVES W/IN 24 HRS PRE-CROSS CLAMP	DDR	01-Apr-94		CLINICAL INFORMATION
24	ARGININE_DON	DECEASED DONOR-WAS DONOR GIVEN ARGININE VASOPRESSIN WITHIN 24 HRS PRE CROSS CLAMP?	DDR	30-Jun-04		CLINICAL INFORMATION
25	ARTIFICIAL_LI_TCR	RECIPIENT ON ARTIFICIAL LIVER AT LISTING	TCR	01-Oct-87		CLINICAL INFORMATION
26	ARTIFICIAL_LI_TRR	RECIPIENT ON ARTIFICIAL LIVER AT TRANSPLANT	TRR	01-Oct-87		CLINICAL INFORMATION
27	ASCITES_TCR	RECIPIENT ASCITES @ REGISTRATION	TCR	01-Apr-94	30-Jun-04	CLINICAL INFORMATION - LIVER MEDICAL FACTORS
28	ASCITES_TRR_OLD	TRR ASCITES	TRR	01-Apr-94	30-Jun-04	TRANSPLANT CLINICAL INFORMATION - RISK FACTORS
29	ASCITES_TX	RECIPIENT ASCITES @ TRANSPLANT	WAITING LIST DATA	30-Jun-04		WAITING LIST DATA
30	BACT_PERIT_TCR	RECIPIENT SPONTANEOUS BACTERIAL PERITONITIS @ REGISTRATION	TCR	01-Apr-94		CLINICAL INFORMATION - LIVER MEDICAL FACTORS
31	BACT_PERIT_TRR	RECIPIENT SPONTANEOUS BACTERIAL PERITONITIS @ TRANSPLANT	TRR	01-Apr-94	01-Jan-07	TRANSPLANT CLINICAL INFORMATION - RISK FACTORS

19	1176563	1049375	4	53	53	10	M
20	831454	693881	4	60	60	2	M
21	986374	887822	4	55	55	3	M
22	1025735	921721	4	68	68	10	F
23	1177129	1049195	4	61	61	3	M
24	418468	615897	4	53	52	5	F
25	981983	884410	4	69	68	4	F

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