A2 Data Sources and Biases 🖥️

Welcome to A2! Please enter answers to the questions in the specified Markdown cells below. When you are done with the assignment, export this file as a PDF and submit to Canvas.

Learning Objective 🌱

In this assignment, you will explore some useful sources of healthcare data and structured biomedical knowledge. There is a vaste trove of resources available to you, and it is important to become familiar with digging through documentation to find what you need for a given project so you don't end up re-inventing the wheel.

Resources 📖

- Refer to the slides from Lecture 3 (Health Care Utilization Databases) for examples of databases and their characteristics.
- The Stanford Data Farm provides detail on more than 150 healthcare-related databases available to Stanford investigators, including some from this homework. This is a very useful resource for accessing datasets that you can use in your own research!

1. Exploring Healthcare Data Sources 🏥 (21 points)

Instructions ☑️

For each of the data sources below, please provide the following:

- `1-Sentence Summary`: Provide a brief summary describing what the resources is
- `Unit(s) of Observation`: The entity or element that is being studied, observed, or recorded in the resource
- `Data Element(s)`: (aka Data Item or Data Attribute) is a specific piece of information or characteristic that is being collected, stored, and managed that

describes each unit of observation.

- `Time Span`: The time span that the records in the resource cover
- `Number of Records`: The number of records in the resource.
- `Creator(s) or Curating Institution`: The entity responsible for creating/curating the resource
- `Potential Linkages`: Other resources that could be easily linked or are designed to be used with this resource. Generally these are resources that are easily linked by common identifiers. For example: a PubMedID or patient identifier (like a SSN) may be used to link entries between databases.

A completed example for the BioPortal resource is included below.

`Example:` BIOPORTAL (`0 Points`)

`1-Sentence Summary`

BioPortal is a repository of biomedical ontologies and mappings between ontologies and concepts, which also offers a software service to recommend ontologies and annotate resources with ontology concepts, as well as a resource index of existing annotations.

`Unit(s) of Observation`

ontology, concept, concept-concept mapping, ontology-ontology mapping

`Data Element(s)`

ontology: acronym, visibility, Bioportal PURL, description, status, format, contact, home page, publications page, documentation page, categories, groups, license information, number of classes, number of individuals, number of properties, maximum depth, maximum number of children, average number of children, classes with a single child, classes with more than 25 children, classes with no definition, visits, release date, upload date, projects using ontology. Concept: ID, preferred name, subClassOf, ontology-specific values

`Time Span`

2005-2016

Number of Records

Ontologies: 535; classes: 7,338,810; resources indexed: 48; indexed records: 39,359,542; direct annotations: 95,468,433,792; direct plus expanded annotations: 144,789,582,932.

Creator(s) or Curating Institution

National Center for Biomedical Ontology

Potential Linkages

UMLS terminologies; the 48 resources indexed with concepts from BioPortal; text that contains mentions of concepts in any ontology available through BioPortal

---

1.1: ClinicalTrials.gov (7 points)

1-Sentence Summary (1 point)

ClinicalTrials.gov is a comprehensive online database that provides information about the clinical research studies to the public, researchers, and healthcare professionals.

Unit(s) of Observation (1 point)

Clinical Trial

Data Element(s) (1 point)

The data contains the following elements:

1. Study Details:
    - Brief Summary
    - Detail Description
    - Official Titles
    - Conditions
    - Intervention / Treatment
    - Other Study ID Numbers
    - Obsolete Identifiers

- Contacts and Locations
- Participants Criteria
  - Elibility Criteria
    - Description, Ages Elibile for Study, Sexes Elibile for Study, Accepts Healthy Volunteers
- Study Plan
  - Design Details
    - Primary Purpose, Allocation, Interventional Model, Masking
  - Intervention / Treatment
- Primary Outcome Measures
  - Outcome Measure, Measure Description, Time Frame
- Collaborators and Investigators
  - Sponsor, Collaborators, Investigators
- Study Record Dates:
  - Study Registration Dates
    - First Submitted, First Submitted that met QC Criteria, First Posted
  - Study Record Updates
    - Last Updated Submitted that met QC creteria, Last updated Posted, Last Verified
- Terms related to this study
  - Keywords provided by sponsors, Additional Relevant MeSH Terms
- Study Start
- Primary Completion
- Study Completion
- Enrollment
- Study Type
- Phase

2. Researcher Review:
   - Trial Contacts
     - Contacts
   - Study Record Dates
     - First Submitted, First Posted, Last Update Posted, Last Verified
   - Outcome Measures:
     - Change History, Primary (Current), Primary (Original), Secondary (Current), Secondary (Original), Other Pre-specified (Current), Other Pre-specified (Original)
   - Trial Description
     - Brief Title, Official Title, Brief Summary, Detailed Description, Study Type,

Study Phase, Study Design, Condition, ntervention, Study Arms, Publications
- Recruitment Information:
  - Recruitment Status, Enrollment, Original Enrollment, Study Start Date, Primary Completion Date, Study Completion Date, Eligibility Criteria, Sex/Gender, Ages, Accepts Health Volunteers, Location Countries, Removed Location Countries
- Administrative Information
  - NCT Numbers, Other Study ID Numbers, Has Data Monitoring Committee, US FDA regulated product, IPD Sharing Statement, Current Responsible Party, Original Responsible Party, Current Study Sponsor, Original Study Sponsor, Collaborators, Investigators, PRS Account

3. Result
   - Result Overview
     - Recruitment Status, Primary Completion Date, Study Completion Date
4. Record History:
   - Version, Date Submitted, Changes

`Time Span` (`1 point`)

The service launched on 2000 and still working till present. In terms of data, they collected the data from September 1997 to Present. The earliest submitted study recored in the system was 1999-09-17 (First submitted date)

`Number of Records` (`1 point`)

As of 2024-10-02, they have total of 510,898 studies stored in their database

`Creator(s) or Curating Institution` (`1 point`)

ClinicalTrials.gov is under National Library of Medicine. Its main data source are the sponsors or the investigators who submit and update the information about studies voluntarily

`Potential Linkages` (`1 point`)

Registry deposition and Publication, Drug Bank, SNOMED

1.2: FDA Adverse Event Reporting System (FAERS) (7 points)

1-Sentence Summary (1 point)

FAERS is an FDA-maintained database of adverse events, medication errors, and quality complaints related to drugs and biologics, used for post-marketing safety surveillance and coded using international standards.

Unit(s) of Observation (1 point)

Adverse Event, Safety Report

Data Element(s) (1 point)

The data include the following:

- We can see the following data item in the reports: safetyreportversion, safetyreportid, primarysourcecountry, occurcountry, transmissiondateformat, transmissiondate, reporttype, serious, seriousnessdeath, seriousnesslifethreatening, seriousnesshospitalization, seriousnessdisabling, seriousnesscongenitalanomali, seriousnessother, receivedateformat, receivedate, receiptdateformat, receiptdate, fulfillexpeditecriteria, companynumb, duplicate, reportduplicate, duplicatesource, duplicatenumb, primarysource, reportercountry, qualification, sender, sendertype, senderorganization, receiver, receivertype, receiverorganization, patient, patientonsetage, patientonsetageunit, patientagegroup, patientsex, reaction, reactionmeddraversionpt, reactionmeddrapt, reactionoutcome, drug, drugcharacterization, medicinalproduct, drugbatchnumb, drugauthorizationnumb, drugstructuredosagenumb, drugstructuredosageunit, drugseparatedosagenumb, drugintervaldosageunitnumb, drugintervaldosagedefinition, drugdosagetext, drugdosageform, drugadministrationroute, drugindication, drugstartdateformat, drugstartdate, actiondrug, drugadditional, activesubstance, activesubstancename, summary, narrativeincludeclinical
- In gener we have the following:
    - demographic and administrative information and the initial report image ID number (if available)
    - drug information from the case reports

- reaction information from the reports
- patient outcome information from the reports
- information on the source of the reports
- a "README" file containing a description of the files

`Time Span` (`1 point`)

2012Q4 - Present

`Number of Records` (`1 point`)

As of 2024-10-02, total of 29,153,222 records are available in the database

	Total Reports	Expedieted	Non-Expedited	Direct	BSR
Total Reports	29,153,222	15,957,059	11,911,258	1,284,042	863

`Creator(s) or Curating Institution` (`1 point`)

U.S. Food & Drug Administration

`Potential Linkages` (`1 point`)

Drug bank, UMLS, SNOMED,

`1.3:` National Inpatient Sample (NIS) (`7 points`) - Done

`1-Sentence Summary` (`1 point`)

The NIS is a comprehensive, publicly accessible healthcare database that provides national estimates and insights on various aspect of inpatient care in US based on the data from millions of hospital.

`Unit(s) of Observation` (`1 point`)

Patient, Diagnosis, Diseases, Treatments

Data Element(s) (1 point)

NIS Data contains clinical and nonclinical data element for each hospital stay including:

- International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis, procedure, and external cause of injury codes prior to October 1, 2015
- International Classification of Diseases, Tenth Revision, Clinical Modification/Procedure Coding System (ICD-10-CM/PCS) diagnosis, procedures, and external cause of morbidity codes beginning October 1, 2015
- Patient demographic characteristics (e.g., sex, age, race, median household income for ZIP Code)
- Hospital characteristics (e.g., ownership)
- Expected payment source
- Total charges
- Discharge status
- Length of stay
- Severity and comorbidity measures
- AHRQ software tools (not available for 2016-2017)

Time Span (1 point)

Based on the website, the time span for this service is 1998 - 2021

Data Name	Time Span
NIS Core	2000, 2002, 2004, 2012 - 2021
NIS DX_PR_GRPS	2012 - 2021
NIS Hospital	2012 - 2021
NIS Severity	2012 - 2021

Number of Records (1 point)

Based on the official website: Unweighted, NIS contains data from around 7 million hospital stays each year. Weighted, it estimates around 35 million hospitalizations nationally. Here are the breakdown collected from Stanford PHS

Data Name	Count
NIS Core	70,264,286

NIS DX_PR_GRPS	55,969,502
NIS Hospital	45,140
NIS Severity	58,096,244

source: https://redivis.com/datasets/2g37-7mghj1wyb/tables

`Creator(s) or Curating Institution` (`1 point`)

Agency for Healthcare Research and Quality (AHRQ)

`Potential Linkages` (`1 point`)

ICD Code, Drug Bank

---

2. Databases and Schemas 💿 (19 points) - Done

Instructions ☑️

The purpose of these questions is to broaden your understanding of databases and how they are organized. For each question, read the prompt and fill out the requested specific information of interest.

`2.1:` (`11 points`)

> You are using MIMIC data. While browsing the `NOTEEVENTS` table, you find a note that contains an interesting hypothesis about the cause of diabetic symptoms in the patient it describes. You wonder if the doctor who wrote the note has experience treating diabetic patients.

- In simple english, describe how you could find out if the doctor who wrote that note has ever previously written a note about a patient with an pre-existing ICD9 diagnosis for diabetes?
- Please explicitly mention which tables, features, and any time filtering you would use to achieve this task. We expect specific details for full credit.

Here are the tables I will use and details steps I will do to check if the doctor who wrote the note has experience treating diabetic patients:

- Tables needed:
  - NOTEEVENTS
  - D_ICD_DIAGNOSES
  - DIAGNOSES_ICD
  - ADMISSIONS
- Steps:
  - Step 1: First, find out all the `icd9_code` related to diabetic from `D_ICD_DIAGNOSES.csv`. This will give us all codes related to diabetic.
  - Step 2: Identify doctor's `CGID` who wrote the note, then from `NOTEVENETS` we can filter out all notes written by the doctor using doctor's `CGID`. This can also give us the subject_id (patient) he had treated.
  - Step 3: Then using the `subject_id` we retrieved from Step 2 and the `icd9_code` from Step 1, we can filter out all patient treated by the doctor we interested with diabetic
  - Step 4: Join the step 3 result with `ADMISSION` table to filter out the patient who has pre-existing diabetics condition prior the date when the notes was written

2.2: (8 points)

> You have access to a large database consisting of three tables of data on patients:

- demographics (`person_ID` , `date_of_birth`),
- visit history (`person_ID` , `date_of_visit` , `provider_seen`)
- drug prescription history (`person_ID` , `drug_prescribed` , `date_of_prescription`).

You are interested in finding all patients of ages 18 to 50 with a prescription for a short-acting stimulant.

You have narrowed your list down to 10 stimulants but, upon skimming through the drug prescription table, you find multiple variations of each stimulant. For example, for the drug ingredient ‘Focalin’ you find the following variations:

drug_name

Focalin XR
Focalin XR 10 mg oral capsule, extended release
Focalin XR 15 mg oral capsule, extended release
Focalin XR 20 mg oral capsule, extended release
Focalin XR 20mg
Focalin XR 30 mg oral capsule, extended release
Focalin XR 35 mg oral capsule, extended release
Focalin XR 5 mg oral capsule, extended release

The table is very long and you do not have time or expertise to look through it and find all of the variations for each drug.

You realize you are missing a table in your database that would allow you to answer your question.

- Describe the columns that this missing table should have and the relationship between them (hint: it has two columns).
- What columns would this table have in common with the other tables in the database? (10 points)

We can call the missing table as `drug` and it will contain 2 columns: `ingredients` and `variations` as follow:

ingredients	variation
Focalin	Focalin XR|Focalin XR 10 mg oral capsule, extended release|Focalin XR 15 mg oral capsule, extended release| Focalin XR 20 mg oral capsule, extended release| Focalin XR 20mg| Focalin XR 30 mg oral capsule, extended release |Focalin XR 35 mg oral capsule, extended release| Focalin XR 5 mg oral capsule, extended release
Ingredient 2	medicine 1|Medicine 2...

With this new table, we can create a query to get all patient of ages 18 to 50 with particular ingredients. For example:

```
WITH split_variations AS (
    SELECT
        ingredients,
        TRIM(value) AS single_variation
    FROM
```

```
            drug
        CROSS APPLY STRING_SPLIT(variation, '|')
    ),
    patient_age AS (
        SELECT
            person_ID,
            DATEDIFF(YEAR, date_of_birth, GETDATE()) -
            CASE
                WHEN DATEADD(YEAR, DATEDIFF(YEAR,
date_of_birth, GETDATE()), date_of_birth) > GETDATE()
                THEN 1
                ELSE 0
            END AS age
        FROM
            demographics
    )
    sv.ingredients
    FROM
        patient_age pa
    JOIN
        drug_prescription_history dph ON pa.person_ID =
dph.person_ID
    JOIN
        split_variations sv ON dph.drug_prescribed =
sv.single_variation
    WHERE
        pa.age BETWEEN 18 AND 50
        AND sv.ingredients IN ('Focalin', 'OtherIngredient1',
'OtherIngredient2')
    ORDER BY
        pa.person_ID, sv.ingredients, dph.drug_prescribed;
```

In this new table, the `variant` will contains the name of the drug (variant) which can be found in `drug_prescribed`

---

3. How data (and biases) are born (`40 points`)

Instructions ☑️

This question should get you thinking about where and how different kinds of data

are generated, and how that also generates the biases that come with them. Read the following prompts and answer the questions in the specified locations.

3.1: (8 points)

> A healthcare database has a field `IS_SMOKER` for each patient.

- How do you believe this information would be measured and put into the database?

There are several ways to measured and recorded into the database:

1. Self-reporting - One of the most common method is asking the patient directly during the intake, medical visits like routine check-ups or using survey. Patients are typically asked whether they smoke, have they smoked in the past or have never somke. This kind of information would been entered into the database by healthcare staff. This method could introduce potential bias because of the following:
   - Underreporting - Patients might not disclose their smoking habits accurately due to different kind of reasons like religion, culture or even the society norms
   - Misclassification - Patients might not be able to classified themselve correctly if the categories are too broad or not clearly defined. For example, "Occational Smoker" vs. "Dialy Smoker")
   - Memory recall: Patient may not remeber their smoking history correctlt, especially if they quit smoking years ago.
2. Clinician Reported - Healthcare providers may record smoking status based on the following: a) Clinical obsewrvation during the consultation b) Electronic Health records c) Medical history review However, these methods can also introduce bias:
   - Subjectivity - Clinician might make assumptions about patient7s smoking habit based on physical appearance or associated health conditions which potentially leading to misclassification
   - Incomplete Data - If clinician forget to ask the question during client visit, the information not not be updated or accurately recorded over the time

3.2: (8 points)

> You want to do an analysis looking at the prevelence of smoking, and you are thinking about utilizing the `IS_SMOKER` field to identify patients that actively smoke.

- What are the factors that might cause either over-reporting of smoking (more people are marked as smokers in the database than there actually are) or under-reporting of smoking (the opposite)?

- Do you think it is more likely that smoking is over or under-reported? Provide a brief explanation of your thinking.

Here are some key factors contributing to over-reporting or under-reporting:

- Factors contribute to over-reporting:
  1. Outdated Information - If patient's record is not updated after they quit smoking, they will still be considered as smoker, leading to over-reporting
  2. Data Entry Error - Clinician can make mistakes when entering the data. For example, they may encode non-smoker as smoker which leading to over counting
- Factors contributed to under-reporting:
  1. Data Entry Error - This is an opposite case from over-reporting that clinician mistakenly encodes smoker as non-smokers which leads to underreporting
  2. Incomplete record - Not all patient may be asked about their smoking status in their every healthcare interaction. This leads to the possibility of missing or incomplete data
  3. Social Stigma

In my opinion, it is most likely that smoking is under-reported in most of the datasets due to the following reasons:

1. Social norms - The patient may under-report their smoking status when smoking is consider socially undesirable
2. Incomplete or outdated data - For example, a patient quick smoking might not be categorized correctly especially when the answer of `IS_SMOKER` is binary (YES/NO) which is not able to capture the past behavior.
3. In the healthcare setting, some patients will underreport their smoking status because they want to present themselves in a healthier light or to avoid judgement from medical professionals

`3.3: (8 points)`

> The data you are utilizing for the analysis has patients from many different backgrounds.

- Can you think of any patient populations that might have more under- or over-reporting of their smoking than others? Please briefly explain your reasoning.

Here are some potential patient populations that might under- or over-reporting of smoking:

- Under-Reporting:
  - Teenagers/Adolescents - This group of people might underreport to avoid parental disapproval or legal consequences. For example, some states impose penalty for Minors possessing or purchasing tabaco products
  - Pregnant women - Expecting mother might underreport due to the social norms/stigma and guilt associated with smoking during pregnancy period.
  - Patient with specific medical conditions - For example, the patient with lung cancer or heart disease might underreport their smoking status due to felling of shame or fear of doctor's / society's judgement
- Over-Reporting:
  - Patient wants to quit smoking might overreport their smoking status so they will be qualify for smoking cessation programs or other related benefits
- Both Under- and Overreporting:
  - Elder patient - This group of people might under- or over-reporting due to memory issue or due to the different cultural norms from their childhood

`3.4: (8 points)`

> You calculate the correlation between smoking and year of birth between 1980 and 2010.

- What do you think the relationship between these two variables is?
- Based on your answers above, do you think the true correlation is stronger or weaker than the what you calculated?

Answer:

- The relationship between these two variables (`smoking` and `year of birth`), for the time range 1980 - 2010, is `NEGATIVE CORRELATION` . with the following reasons:
  - Increased awareness - Over this period of time (1980 - 2010), people are more aware with the health risk associated with smoking due to aggressive Anti-smoking campaigns and education programs

    - Regulations - Strict regulation on selling tobacco products (specially to minors), advertising, and smoking in public places making smoking more difficult
    - Price increases - Governments have raised tax on tobacco products, making smoking become more and more expensive
    - Social norms/Cultural shift - Smoking become less socially acceptable over the time
- In my opinion, the true correlation might be stronger than what I am calculated - That means stronger `NEGATIVE CORRELATION`. It might mainly due to the people under-report their smoking status.

3.5: (8 points)

> Sometimes a data element of interest does not appear in the data you have access to. A possible work-around is to use a proxy data element known to be correlated with the data element of interest. For example, Frankovich et al. used aspirin records as a proxy for antiphospholipid antibody labs.

- What are some strategies you might use to assess or choose a proxy for a variable of interest? (any variable, not just the `IS_SMOKER` field mentioned in previous question parts)

In this case, I will be consider the following strategies to choose possible proxy:

1. Subject Matter Expert - Consult with Subject Matter Expert can help us identify possible variables that are theoretically or clinically related to the variable of interest
2. Literature review - Experience form the past experiment can be a good source to help us to identify the possible proxies for similar variables in comparable context
3. Domain knowledge - For example, if the data `IS_SMOKER` is not in the dataset collected, we can try to look for other proxies like lung's x-ray image, any lung's related medication... etc. However, this needs some domain knowledge to help us idetify the right proxies
4. Sensitivity Analysis - This analysis can help us understand the relationship between the input (proxy) and the output (variable of interest)
