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 🏥 (28 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 public registry and results database of privately and publicly supported clinical studies in humans conducted worldwide.

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

Study record or trial identified by an unique NCTID.

`Data Element(s)` (`1 point`)

NCT ID; study title; conditions; interventions; sponsor; study type and phase; design; enrollment; locations; eligibility; key dates; recruitment status; outcome measures; adverse events and results summaries.

`Time Span` (`1 point`)

2000–present (first postings in February of 2000).

`Number of Records` (`1 point`)

Over 480,000 registered study records worldwide as of April 2025.

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

National Library of Medicine (NLM), National Institutes of Health (NIH).

`Potential Linkages` (`1 point`)

PubMed articles via NCT IDs; FDA records/labels; RxNorm/DrugBank for medical interventions; trial sponsor identifiers;

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

1–Sentence Summary (1 point)

FAERS is the FDA's postmarketing safety database of adverse event and medication error reports for drugs and biologics.

Unit(s) of Observation (1 point)

Individual case safety report (ICSR), adverse events, and medication errors, identified by CASEID and ISR.

Data Element(s) (1 point)

CASEID; ISR; receipt and report dates; patient age, sex, weight; reporter type and country; suspect and concomitant drugs; indication; reaction terms (MedDRA PT); outcomes/seriousness; manufacturer number; primary source.

Time Span (1 point)

2004–present for public quarterly FAERS data. FAERS Quarterly Data Extracts (QDEs) are available starting from the first quarter of 2004.

Number of Records (1 point)

Over 31.8 million cumulative ICSRs, adverse event case reports, as of April, 2025.

Creator(s) or Curating Institution (1 point)

U.S. Food and Drug Administration (FDA), including Center for Dug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER).

Potential Linkages (1 point)

RxNorm/UNII for drug ingredients; National Drug Code (NDC) or Structured Product Labeling (SPL) for products; OpenFDA FAERS API; PubMed via drug/adverse event terms, clinicaltrials.gov, Medical Dicionary for Regulartory Activities (MEDRA).

1.3: Medical Information Mart for Intensive Care (MIMIC) (7 points)

1-Sentence Summary (1 point)

MIMIC is a freely accessible, de-identified database of critical care patients from Beth Israel Deaconess Medical Center, containing detailed clinical data across ICU stays.

Unit(s) of Observation (1 point)

Patient, hospital admission, and ICU stay records.

Data Element(s) (1 point)

Demographics; admissions and ICU stay details; vital signs; laboratory results; medications and infusions; procedures; diagnoses (ICD-9/10); microbiology; charted events; imaging and echo reports; notes (discharge, nursing, radiology).

Time Span (1 point)

There are a few different interations of MIMIC, each with a different time range. MIMIC-II: 2001-2008; MIMIC-III: 2001–2012; MIMIC-IV: 2008–2019.

Number of Records (1 point)

MIMIC-III: ~53k distinct hospital admissions and ~61k ICU stays; MIMIC-IV: >70k ICU stays and >380k hospital admissions.

Creator(s) or Curating Institution (1 point)

MIT Laboratory for Computational Physiology (with Beth Israel Deaconess Medical Center).

Potential Linkages (1 point)

Massachusetts death registry for mortality; ICD-9/10, LOINC, RxNorm mappings; OMOP CDM conversions;

1.4: National Inpatient Sample (NIS) (7 points)

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

The NIS is the largest publicly available all-payer inpatient healthcare database in the U.S., a nationally representative sample of hospital stays.

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

Inpatient discharge (hospital stay) records; hospitals (for sampling strata/weights).

`Data Element(s)` (`1 point`)

Patient demographics (age, sex); diagnoses and procedures (ICD-9-CM/ICD-10-CM/PCS); length of stay; charges/costs; payer; admission/discharge status; hospital characteristics; weight variables for national estimates.

`Time Span` (`1 point`)

1988–present with redesigns inbetween. The National Inpatient Sample (NIS), created in 1988 by AHRQ's HCUP, is the largest all-payer inpatient database in the U.S. It was redesigned in 2012 to sample discharges rather than hospitals, improving representativeness and precision. In October 2015, the NIS transitioned from ICD-9-CM to ICD-10-CM/PCS coding, making 2016 the first full year of ICD-10 data.

`Number of Records` (`1 point`)

~7 million unweighted inpatient discharges per year in recent releases, representing over 35 million U.S. hospital stays after weighting.

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

Agency for Healthcare Research and Quality (AHRQ), Healthcare Cost and Utilization Project (HCUP).

`Potential Linkages` (`1 point`)

ICD-9/10 code sets; AHA Annual Survey (hospital characteristics); Census and ACS for area-level SES; CMS provider/hospital identifiers; State Inpatient Databases (SID) and other HCUP datasets.

---

2. Databases and Schemas 💿 (32 points)

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: (22 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.

Starting from the NOTEVENTS table, we notice an interesting hypothesis about the cause of the diabetic symptoms. In order to investigate further, we want to found out if the doctor has ever written a prior note for ICD9 diabetes diagnosis. To find this information, we can traverse MIMIC tables in the following manner :

From the NOTEVENTS table, we have the hospitals admissions id (hadm_id). We can find this row in the DATETIMEEVENTS dataset. From here, we find the corespoding care giver id (cgid). Now that we have identified the doctor, we can go to the CHARTEVENTS dataset. CHARTEVENTs also has hadm_id so could have gone straight here. From CHARTEVENTS, we can get mappings for a doctor or cgid to all the subject_ids that they have treated. So, for the doctor that treated the entry, we can look at all patients that they have treated. Finally, we can traverse to the Diagnoses_ICD table, which can map subject_id to icd9_code. I would join D_ICD_Diagnoses on icd9_code to this dataset and then filter down for everything that had diabetes as the title. We could also do a pubmed / google search to find all ICD codes relating to diabetes. Thus, we have all of the subjects the doctor treated / saw concerning diabetes. We can take this set of patients and go back to NOTEEVENTS and look at all the other subject_ids that the doctor wrote entries for as desired.

`2.2: (10 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)

Missing table: a two-column drug synonym-to-ingredient dictionary many to one mapping. This would allow you to look at the drugs have a standardized list of what is actually included in those drugs. It would relate back to the drug prescription history

table..

- Column 1: `drug_prescribed` (raw string exactly as it appears in prescriptions history table).
- Column 2: `ingredient_standardized` (key active ingredient or RxNorm ingredient code).

Relationship: many variants (brand/dose/formulation/spelling) map to one standardized ingredient. This lets you filter by your list of 10 short-acting stimulant ingredients regardless of naming variations.

Columns in common with existing tables:

- Shares `drug_prescribed` with the drug prescription history table (`person_ID` , `drug_prescribed` , `date_of_prescription`).
- No shared columns with demographics or visit history (you'll join those via `person_ID` from the prescription table when applying age and visit filters).

---

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?

I believe that this field would be measured as a boolean. The 'is' aspect implies taht there is some binary nature to if they are a smoker or not. As for measurements, this field is most likely filled out by physicians after consultation with patients. Alternatively, theis field could be measured via surveys. You most likely wouldn't diagnose as a smoker via an MRI or x-ray due to these costs. You are more likely to diagnose as a smoker and the preform tests to see the damage

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.

There is reason to think both under and over reporting may occur : First, towards the over-reporting case: clinician selects smoker when any history exists (ex-/former smokers), problem-list carryover or template auto-fill, miscoding/NLP mapping errors, inclusion of nicotine-replacement as “smoker." Second, towards the under-reporting case: social stigma and other factors begetting patient nondisclosure, clinician not asking or failing to update status, defaults to “No” in EHR, ambiguity about occasional e-cig/cigar use, and missing data in brief/urgent encounters.

More likely direction: under-reporting. Patients under-disclose and records go stale; defaults and workflow gaps bias toward “non-smoker,” so prevalence from `IS_SMOKER` is typically an underestimate.

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.

- Under-reporting more likely: adolescents/young adults (stigma, parental presence), pregnant patients (fear of judgment), limited English proficiency/low health literacy (misunderstanding/incomplete documentation and again the 'no' default), occupations where disclosure has consequences (healthcare, aviation, etc.). - Over-reporting more likely: older adults/long-term patients with historic “smoker” labels copied forward, former smokers miscoded as current, and inclusion of vaping/cigar/occasional use as “smoker.” - Net effect: overall bias toward under-reporting due to nondisclosure and stale fields, but certain settings (templates/carryover) can inflate over-reporting in older adults.

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?
- Relationship: negative. Later birth cohorts (closer to 2010) have lower smoking prevalence than earlier cohorts (1980), so smoking decreases as year of birth increases.
- True vs calculated: the calculated correlation is likely more negative than truth; the true correlation is weaker. Older cohorts may be over-labeled as smokers (carryover), while younger cohorts under-disclose or aren't updated, exaggerating the cohort difference.

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)

When you lacks the exact variable, you should first make a clear causal case / theory for why a candidate proxy should track the construct and in what direction (for smoking, nicotine-replacement prescriptions can be a plausible example). You could then validate the proxy wherever a gold standard exists, checking discrimination (ROC/AUC), calibration, sensitivity/specificity etc. and ensure the proxy's timing precedes or coincides with the analysis window to avoid reverse causation. Next, you can examine misclassification — especially whether it is differential across age, sex, race, site, or time —and audits data quality (missingness, coding variability, inter-rater reliability), favoring standardized vocabularies such as RxNorm, LOINC, and ICD. By looking at multiple different proxies and vetting them in the aformentioned manner, you should be able to find proxies for a certain variable of interest.

Feedback (`0 points`)

Please fill out the following feedback form so we can improve the course for future students!
