National Institutes of Health Dataset

Christopher Perry & Abel Sandoval
NCorona-Lyme

Big Pharma is hurting drug innovation.

01 **Are pharmaceutical companies oversaturating some drug markets?**

https://www.washingtonpost.com/news/theworldpost/wp/2018/10/17/pharmaceutical/

02 **Do pharmaceutical companies have higher success rates?**

National Institutes of Health Dataset

- **Notable Tables**: Arm Groups, Clinical Studies Main, Collaborators, Primary Outcomes, Secondary Outcomes, Other Outcomes, & Responsible Parties
- **Primary Key**: nct_number

nih_erd_beam.clinical_studies_main_Beam_DF		
PK	nct_number	string
	org_study_id	string
	secondary_id	string
	official_title	string
	brief_summary	string
	overall_status	string
	enrollment	integer
	enrollment_type	string
	start_date	timestamp
	completion_date	timestamp
	completion_date_type	string
	condition	string
	number_of_arms	integer
	number_of_groups	integer
	phase	string
	study_type	string
	study_design	string
	first_received_date	timestamp
	last_changed_date	timestamp
	verification_date	timestamp
	primary_completion_date	timestamp
	lead_sponsor_agency	string
	lead_sponsor_agency_class	string
	overall_official_full_name	string
	overall_official_role	string
	overall_official_affiliation	string
	serialid	integer

aero_modeled.birds_eye_Beam_DF		
PK, FK	nct_number	string
	sponsor	string
	title	string
	start_year	integer
	start_month	integer
	phase	string
	enrollment	integer
	status	string
	condition	string

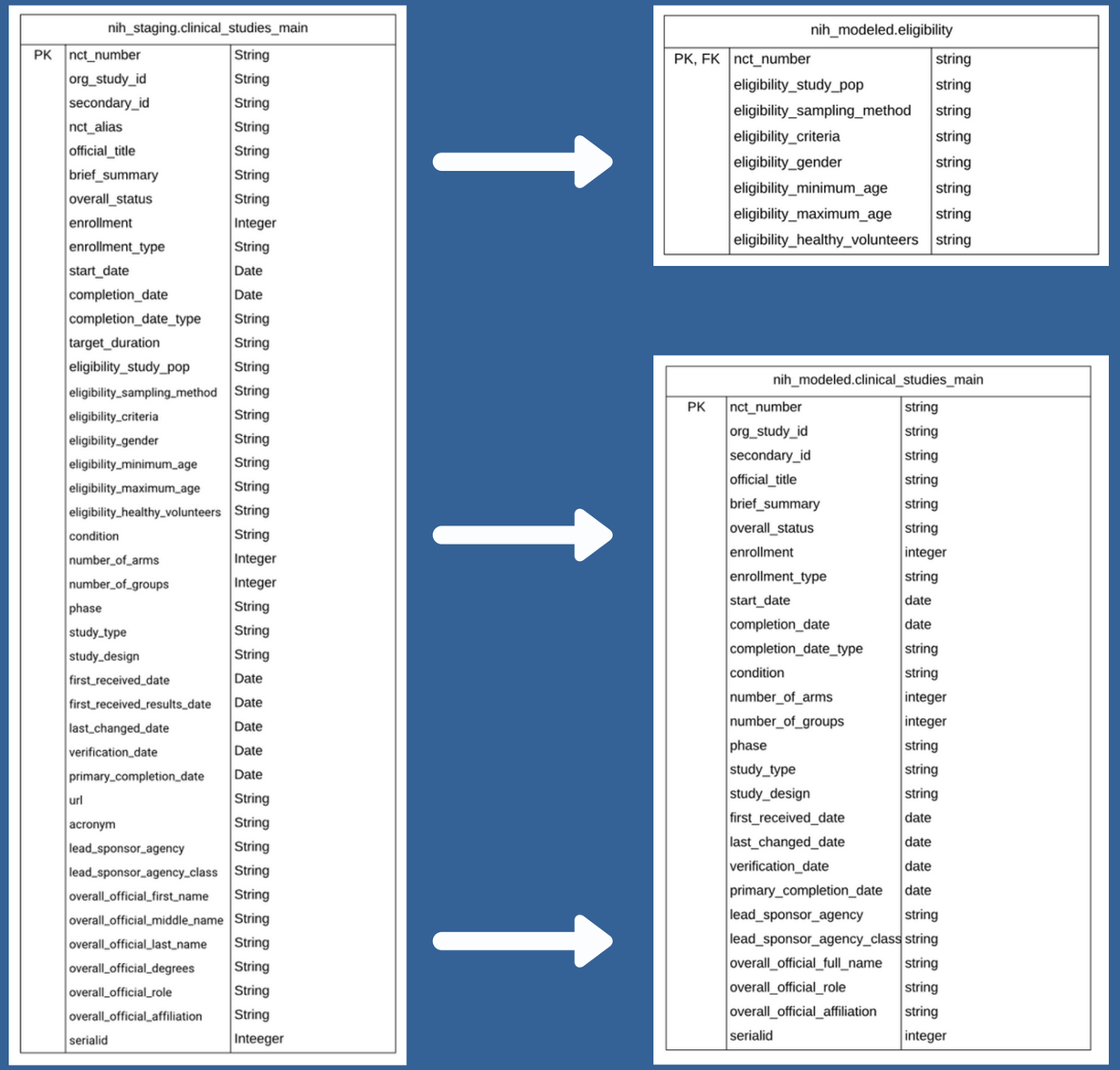
AERO Bird's Eye Dataset

- Notable Features: nct_number, sponsor, title, start_year, start_month, phase, enrollment, status, & condition
- Primary Key: nct_number
- Registered clinical trials from 10 large pharmaceutical companies: AbbVie, Bayer, Gilead, GSK, Johnson & Johnson, Merck, Novartis, Pfizer, Roche, and Sanofi

Modeled Tables

Tasks:

- Remove unnecessary tables
- Remove unnecessary columns
- Cast TIMESTAMP to DATE
- Create Eligibility table



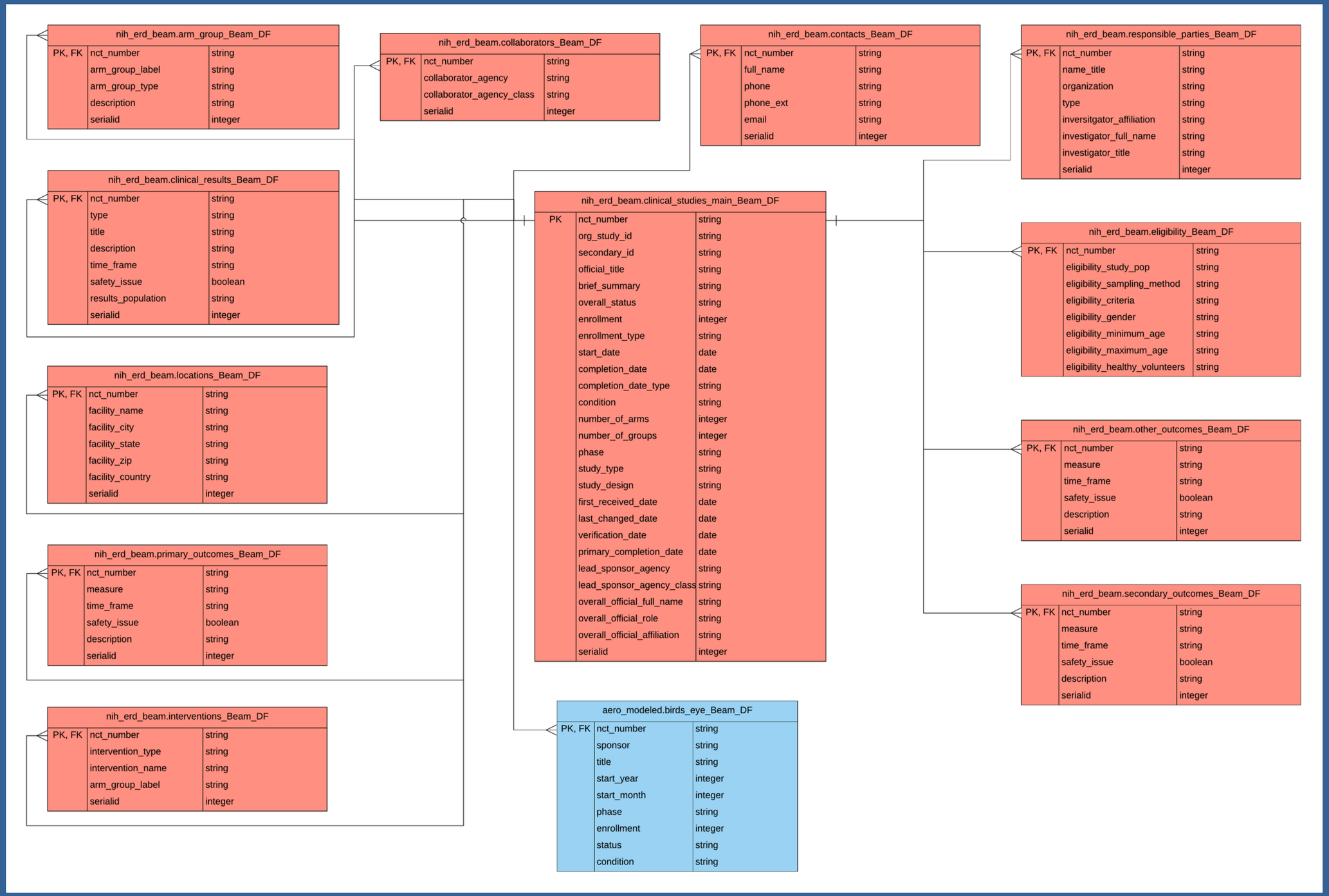
nih_modeled.clinical_studies_main_Beam_DF		
PK	nct_number	string
	org_study_id	string
	secondary_id	string
	official_title	string
	brief_summary	string
	overall_status	string
	enrollment	integer
	enrollment_type	string
	start_date	date
	completion_date	date
	completion_date_type	string
	condition	string
	number_of_arms	integer
	number_of_groups	integer
	phase	string
	study_type	string
	study_design	string
	first_received_date	date
	last_changed_date	date
	verification_date	date
	primary_completion_date	date
	lead_sponsor_agency	string
	lead_sponsor_agency_class	string
	overall_official_full_name	string
	overall_official_role	string
	overall_official_affiliation	string
	serialid	integer

Beam Pipeline

Tasks:

- Make PCollection from data
- Apply PTransform
- Deduplicate Records
- Use GroupByKey()

```
class DedupRecordsFn(beam.DoFn):
  # removes duplicates from table
  def process(self, element):
     nct_number, table_obj = element # table_obj is an _UnwindowedValues object
     table_list = list(table_obj) # cast to list type
     table_record = table_list[0]
     return [table_record]
```


nih_erd_beam.arm_group_Beam_DF
PK, FK nct_number string
arm_group_label string
arm_group_type string
description string
serialid integer
nih_erd_beam.collaborators_Beam_DF
PK, FK nct_number string
collaborator_agency string
collaborator_agency_class string
serialid integer
nih_erd_beam.contacts_Beam_DF
PK, FK nct_number string
full_name string
phone string
phone_ext string
email string
serialid integer
nih_erd_beam.responsible_parties_Beam_DF
PK, FK nct_number string
name_title string
organization string
type string
inversitgator_affiliation string
investigator_full_name string
investigator_title string
serialid integer
nih_erd_beam.clinical_results_Beam_DF
PK, FK nct_number string
type string
title string
description string
time_frame string
safety_issue boolean
results_population string
serialid integer
nih_erd_beam.clinical_studies_main_Beam_DF
PK nct_number string
org_study_id string
secondary_id string
official_title string
brief_summary string
overall_status string
enrollment integer
enrollment_type string
start_date date
completion_date date
completion_date_type string
condition string
number_of_arms integer
number_of_groups integer
phase string
study_type string
study_design string
first_received_date date
last_changed_date date
verification_date date
primary_completion_date date
lead_sponsor_agency string
lead_sponsor_agency_class string
overall_official_full_name string
overall_official_role string
overall_official_affiliation string
serialid integer
nih_erd_beam.eligibility_Beam_DF
PK, FK nct_number string
eligibility_study_pop string
eligibility_sampling_method string
eligibility_criteria string
eligibility_gender string
eligibility_minimum_age string
eligibility_maximum_age string
eligibility_healthy_volunteers string
nih_erd_beam.locations_Beam_DF
PK, FK nct_number string
facility_name string
facility_city string
facility_state string
facility_zip string
facility_country string
serialid integer
nih_erd_beam.other_outcomes_Beam_DF
PK, FK nct_number string
measure string
time_frame string
safety_issue boolean
description string
serialid integer
nih_erd_beam.primary_outcomes_Beam_DF
PK, FK nct_number string
measure string
time_frame string
safety_issue boolean
description string
serialid integer
nih_erd_beam.secondary_outcomes_Beam_DF
PK, FK nct_number string
measure string
time_frame string
safety_issue boolean
description string
serialid integer
nih_erd_beam.interventions_Beam_DF
PK, FK nct_number string
intervention_type string
intervention_name string
arm_group_label string
serialid integer
aero_modeled.birds_eye_Beam_DF
PK, FK nct_number string
sponsor string
title string
start_year integer
start_month integer
phase string
enrollment integer
status string
condition string

Cross-Dataset Queries

```
%%bigquery

select 'clinical studies main' as dataset, MIN(csm.enrollment) as min_enrollment, MAX(csm.enro
from nih_modeled.clinical_studies_main_Beam_DF csm
where csm.lead_sponsor_agency not in (
    select distinct aero.sponsor
    from aero_modeled.birds_eye_Beam_DF aero

    union all

    select distinct csm.lead_sponsor_agency
    from nih_modeled.clinical_studies_main_Beam_DF csm
    where csm.enrollment > 75000
    )

union all

select 'aero bird\'s eye' as dataset, MIN(csm.enrollment) as min_enrollment, MAX(csm.enrollmen
from nih_modeled.clinical_studies_main_Beam_DF csm
where csm.lead_sponsor_agency in (
    select distinct aero.sponsor
    from aero_modeled.birds_eye_Beam_DF aero
    ) and (csm.enrollment <= 75000)
```

```
%%bigquery

(
select 'clinical studies main' as dataset, condition, count(*) as count, ((count(*) / 234703)
from nih_modeled.clinical_studies_main_Beam_DF
where condition in (
    select condition
    from aero_modeled.birds_eye_Beam_DF
    group by condition
    order by count(*) DESC
    limit 10
    )
group by condition
order by count DESC
)

union all

(
select 'aero bird\'s eye' as dataset, condition, count(*) as count, ((count(*) / 13748) * 100)
from aero_modeled.birds_eye_Beam_DF
group by condition
order by count DESC
limit 10
)
```

```
%%bigquery

(
select 'clinical studies main' as dataset, overall_status, count(*) as count, ((count(*) / 234
from nih_modeled.clinical_studies_main_Beam_DF
group by overall_status
having percent > 2
limit 4
)

union all

select 'aero bird\'s eye' as dataset, overall_status, count(*) as count, ((count(*) / 13748) *
from nih_modeled.clinical_studies_main_Beam_DF csm inner join aero_modeled.birds_eye_Beam_DF a
on csm.nct_number = aero.nct_number
where overall_status in (
    select overall_status
    from `probable-pager-266720.nih_modeled.clinical_studies_main_Beam_DF`
    group by overall_status
    having ((count(*) / 234703) * 100) > 2
    limit 4
    )
group by overall_status
```

Airflow DAGs

DAGs: Create Dataset DAGs / Load Tables DAGs / Dummy Operators DAGs
Create Tables DAGs / Dataflow DAGs

```
create_staging >> create_modeled >> branch
branch >> load_arm_groups >> create_arm_groups >> arm_groups
branch >> load_clinical_results >> create_clinical_results >> clinical_results
branch >> load_clinical_studies_main >> create_clinical_studies_main >> create_eligibility >> [clinical_studies_main, eligibility]
branch >> load_collaborators >> create_collaborators >> collaborators
branch >> load_contacts >> create_contacts >> contacts
branch >> load_interventions >> create_interventions >> interventions
branch >> load_locations >> create_locations >> locations
branch >> load_primary_outcomes >> create_primary_outcomes >> primary_outcomes
branch >> load_secondary_outcomes >> create_secondary_outcomes >> secondary_outcomes
branch >> load_other_outcomes >> create_other_outcomes >> other_outcomes
branch >> load_responsible_parties >> create_responsible_parties >> responsible_parties
branch >> load_birds_eye >> create_birds_eye >> birds_eye
```

Future Improvements

01 Analyzing data from other countries that were not included in the data.

02 Further analyze demographics of clinical trials to see if they are generalizable to the population.

Questions?