National Institutes of Health Dataset

Abel Sandoval

Is Big Pharma is hurting drug innovation?

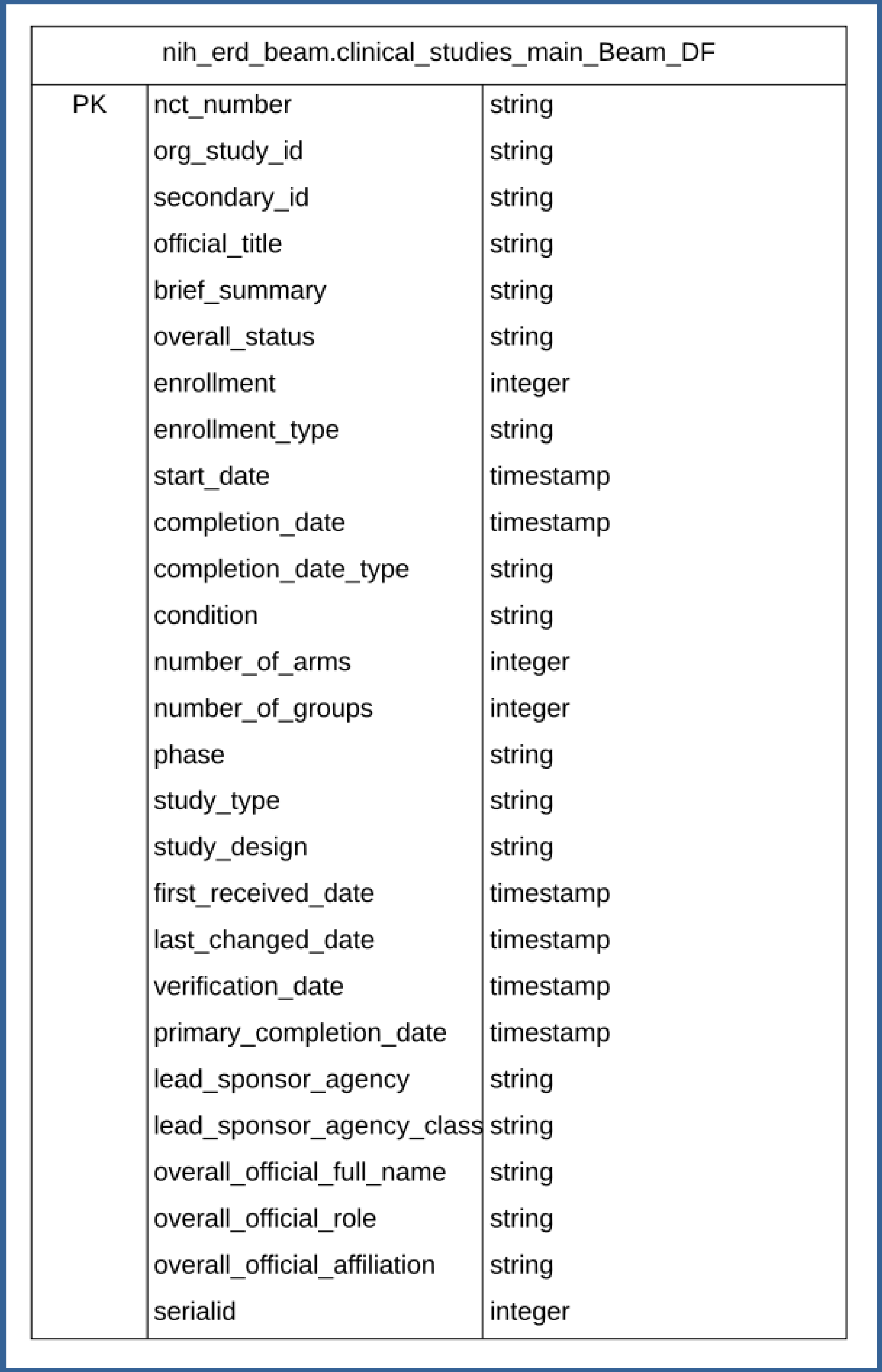
01 Are pharmaceutical companies oversaturating some drug markets?

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

02 Do top pharmaceutical companies have higher success rates?

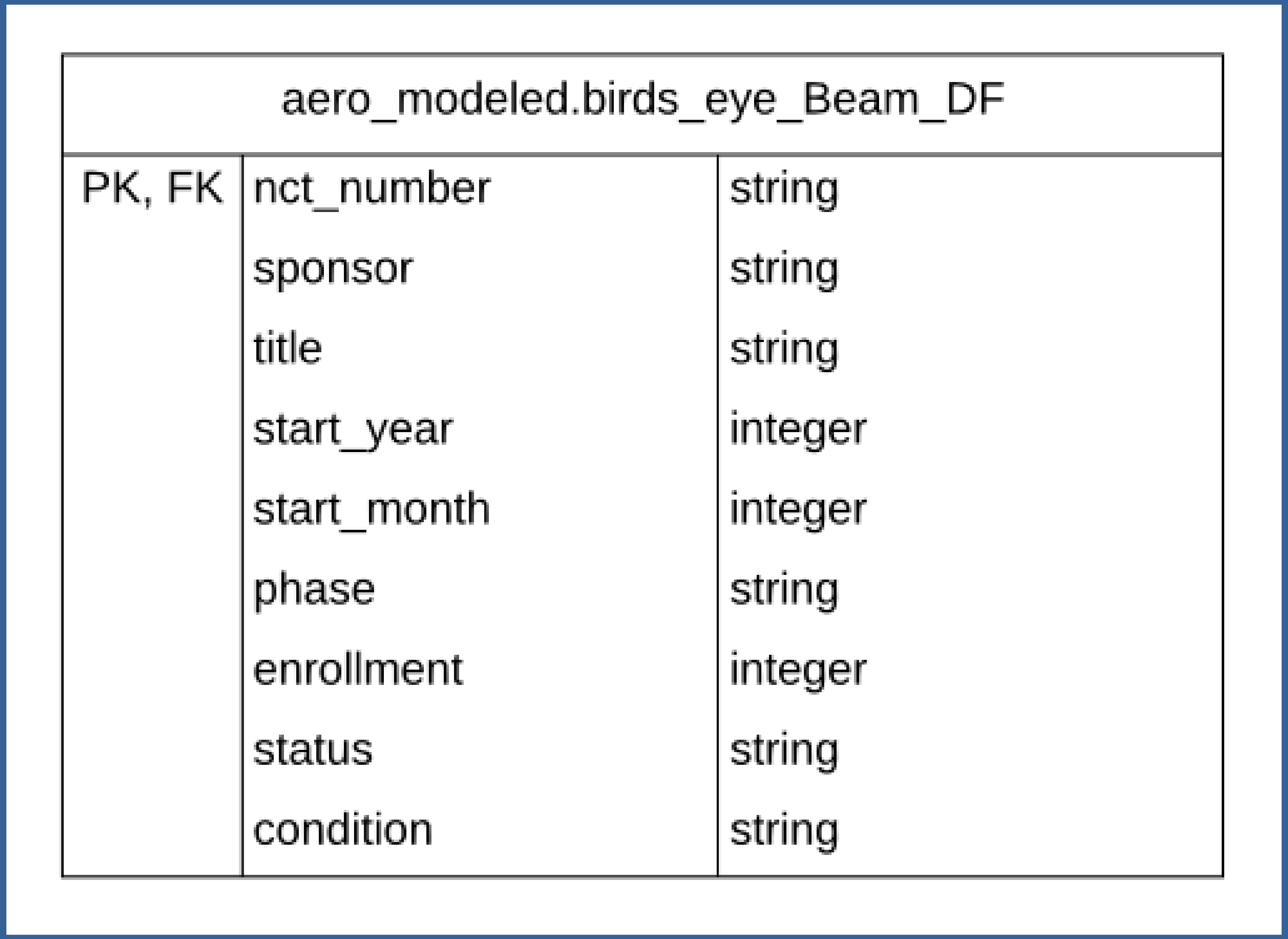
National Institutes of Health Dataset

- 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

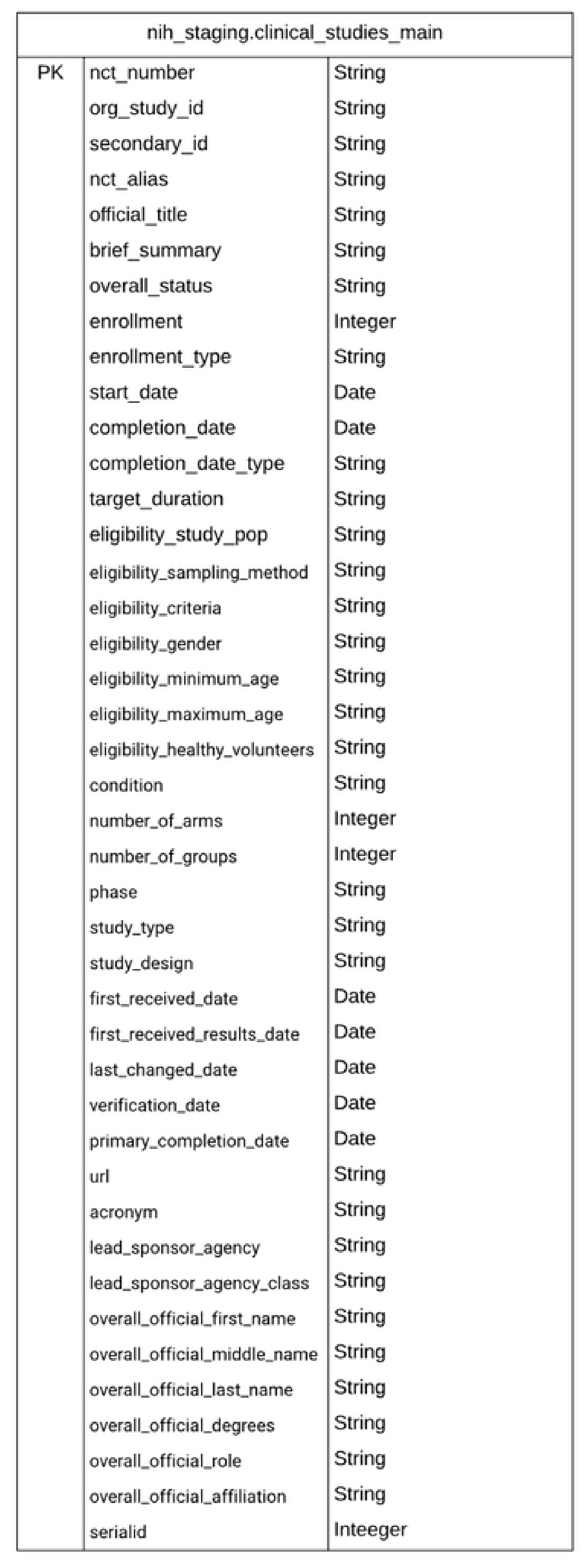
- **Primary Key**: nct_number

- Registered clinical trials from 10 large pharmaceutical companies: AbbVie, Bayer, Gilead, GSK, Johnson & Johnson, Merck, Novartis, Pfizer, Roche, and Sanofi

Modeled Tables

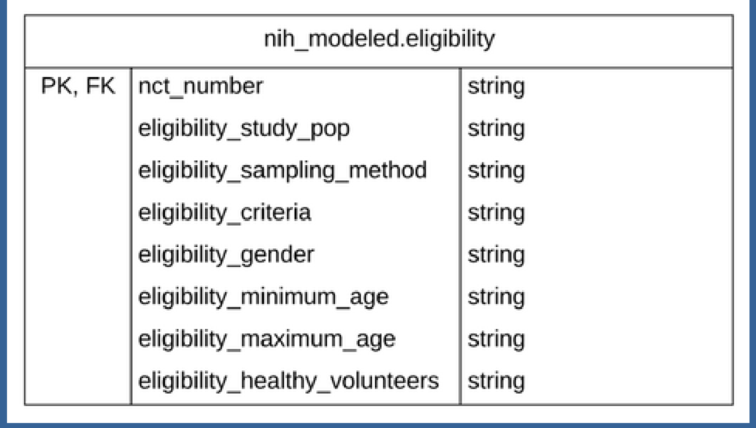
Tasks:

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



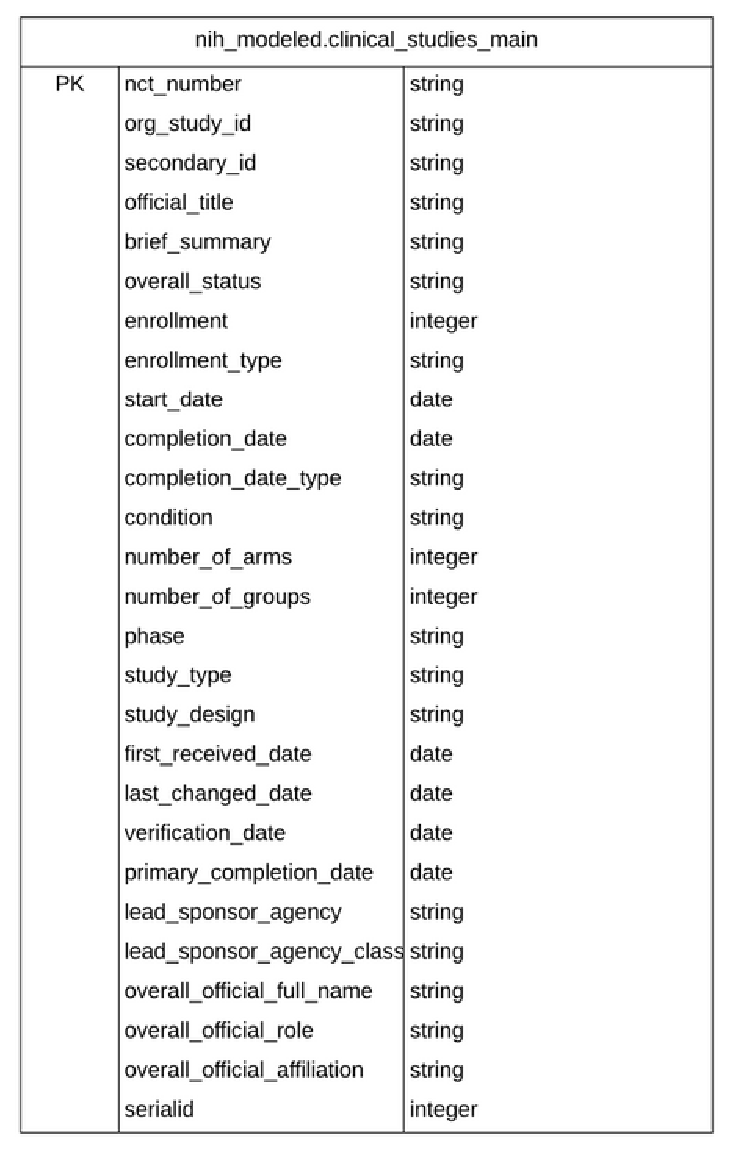
nih_staging.clinical_studies_main

PK	nct_number	String
	org_study_id	String
	secondary_id	String
	nct_alias	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
	target_duration	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
	condition	String
	number_of_arms	Integer
	number_of_groups	Integer
	phase	String
	study_type	String
	study_design	String
	first_received_date	Date
	first_received_results_date	Date
	last_changed_date	Date
	verification_date	Date
	primary_completion_date	Date
	url	String
	acronym	String
	lead_sponsor_agency	String
	lead_sponsor_agency_class	String
	overall_official_first_name	String
	overall_official_middle_name	String
	overall_official_last_name	String
	overall_official_degrees	String
	overall_official_role	String
	overall_official_affiliation	String
	serialid	Integeer



nih_modeled.eligibility

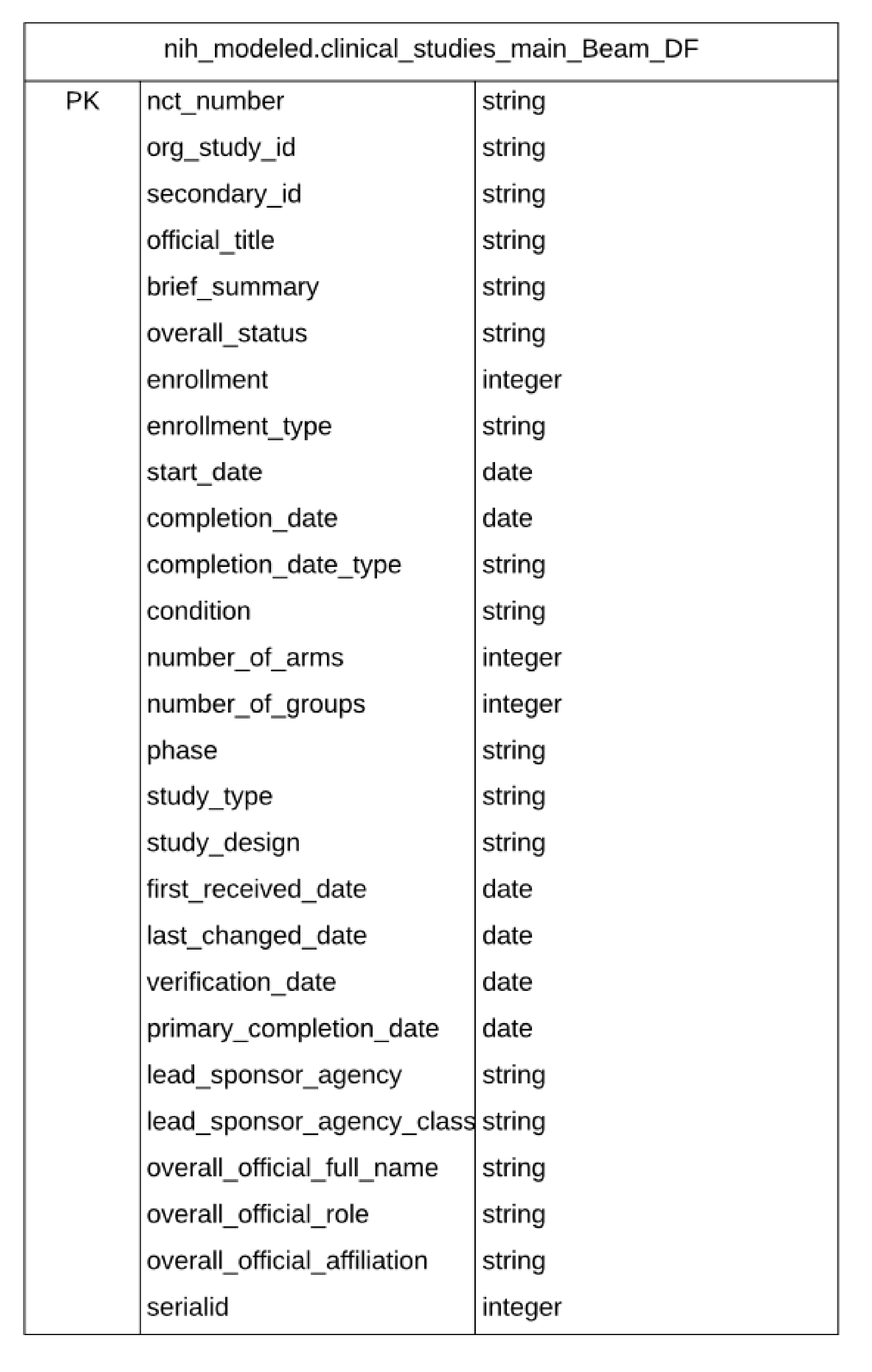
PK, FK	nct_number	string
	eligibility_study_pop	string
	eligibility_sampling_method	string
	eligibility_criteria	string
	eligibility_gender	string
	eligibility_minimum_age	string
	eligibility_maximum_age	string
	eligibility_healthy_volunteers	string



nih_modeled.clinical_studies_main

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

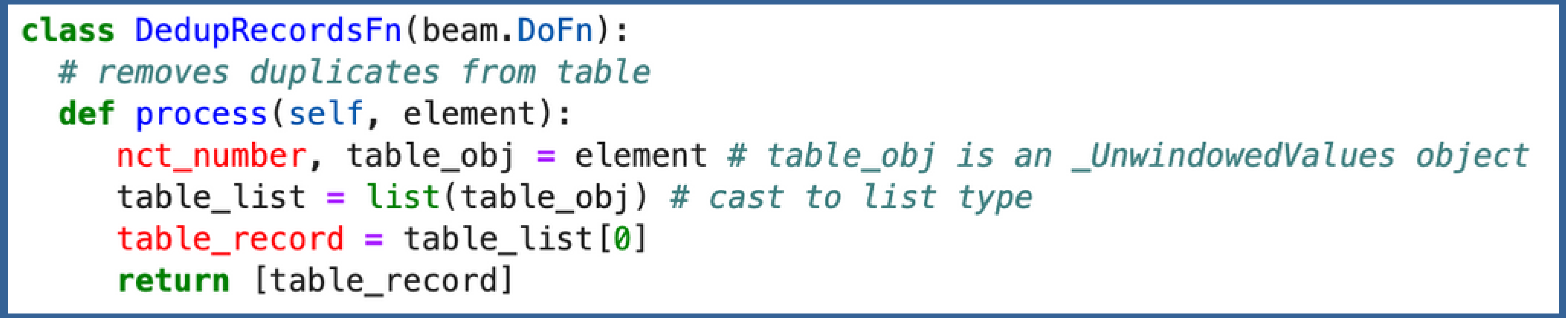


nih_modeled.clinical_studies_main_Beam_DF

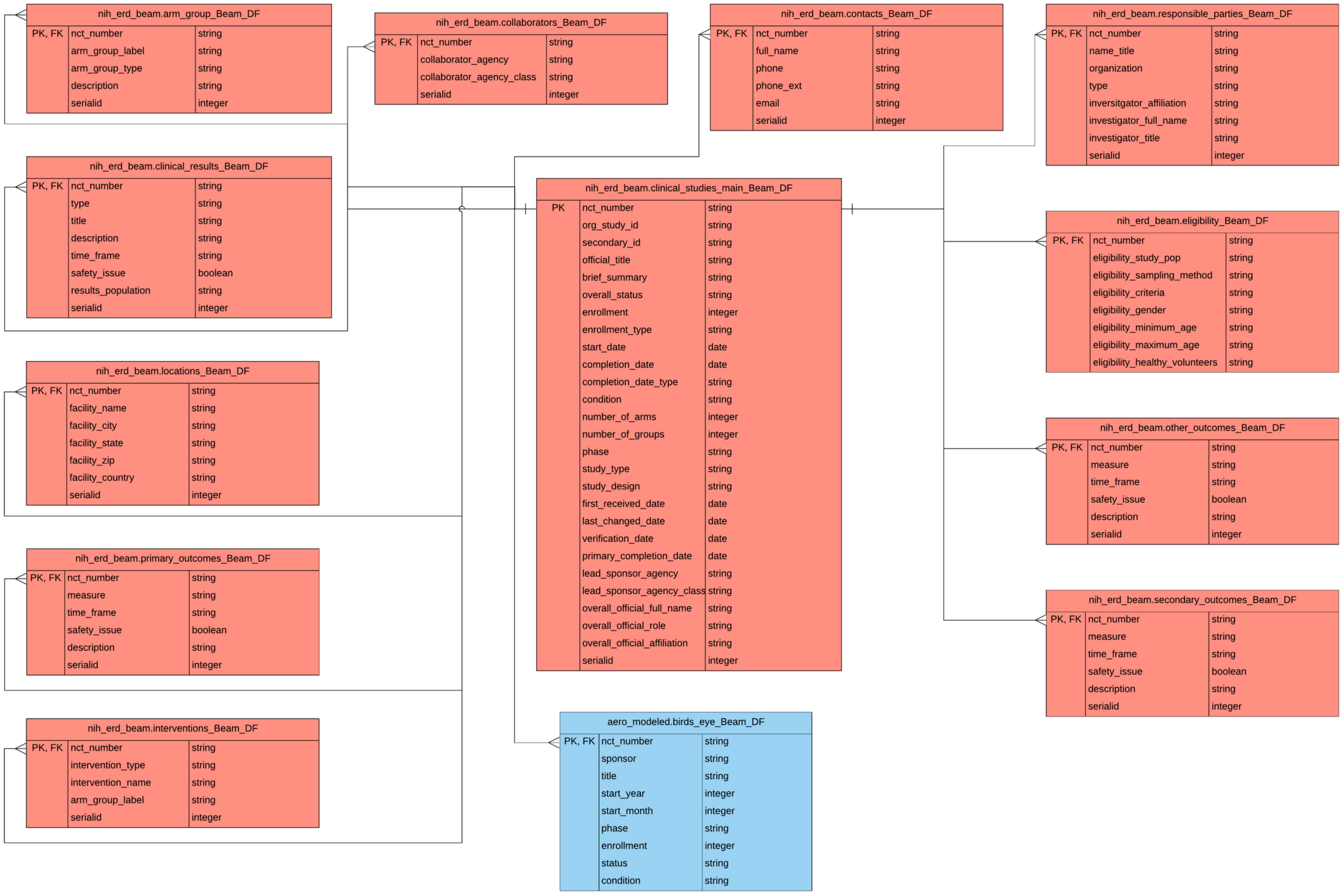
PK	nct_number	string
	org_study_id	string
	secondary_id	string
	official_title	string
	brief_summary	string
	overall_status	string
	enrollment	integer
	enrollment_type	string
	start_date	date
	completion_date	date
	completion_date_type	string
	condition	string
	number_of_arms	integer
	number_of_groups	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

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.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.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.primary_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

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.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

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

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.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

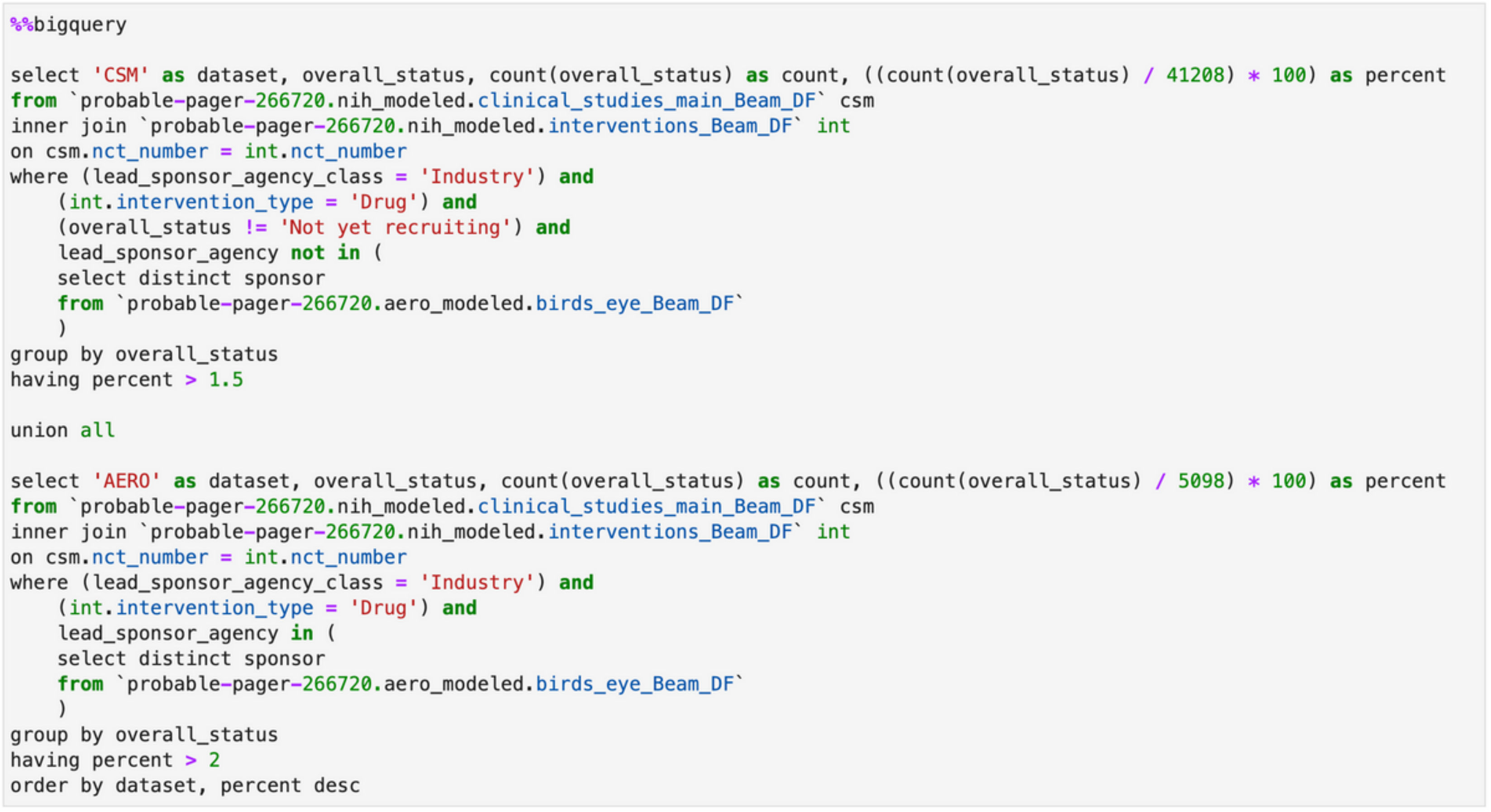
nih_erd_beam.other_outcomes_Beam_DF

PK, FK	nct_number	string
	measure	string
	time_frame	string
	safety_issue	boolean
	description	string
	serialid	integer

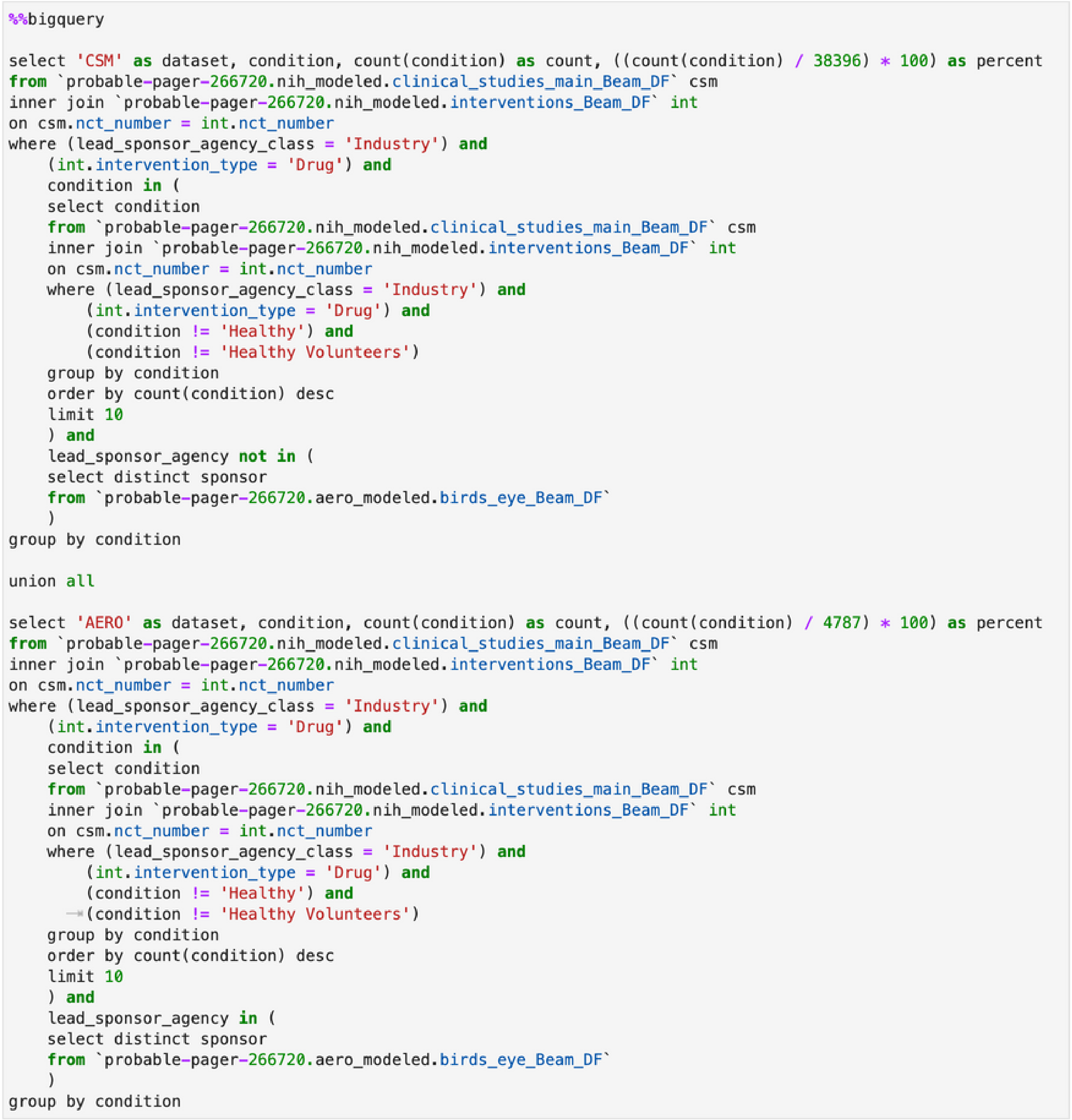
nih_erd_beam.secondary_outcomes_Beam_DF

PK, FK	nct_number	string
	measure	string
	time_frame	string
	safety_issue	boolean
	description	string
	serialid	integer

Cross-Dataset Queries



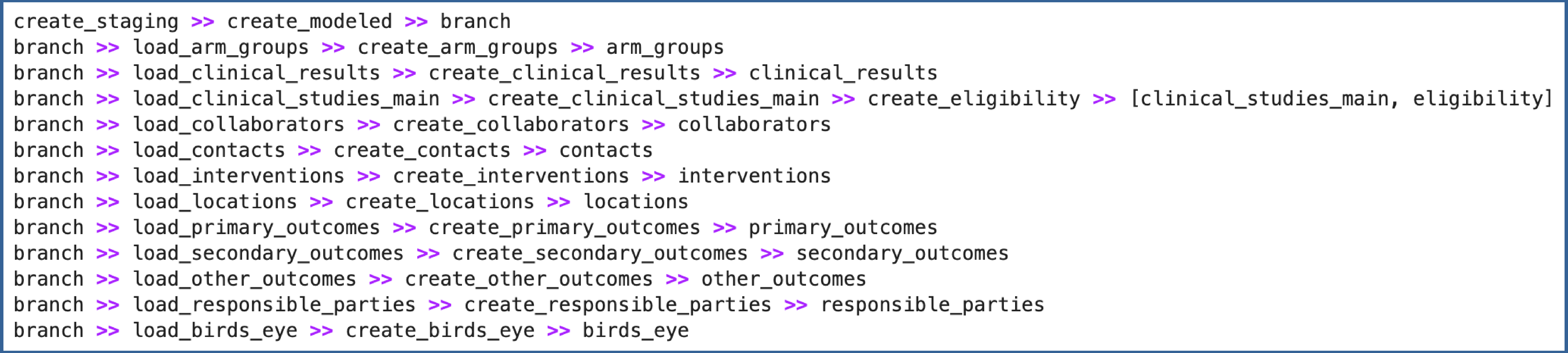
```
%%bigquery

select 'CSM' as dataset, overall_status, count(overall_status) as count, ((count(overall_status) / 41208) * 100) as percent
from `probable-pager-266720.nih_modeled.clinical_studies_main_Beam_DF` csm
inner join `probable-pager-266720.nih_modeled.interventions_Beam_DF` int
on csm.nct_number = int.nct_number
where (lead_sponsor_agency_class = 'Industry') and
    (int.intervention_type = 'Drug') and
    (overall_status != 'Not yet recruiting') and
    lead_sponsor_agency not in (
    select distinct sponsor
    from `probable-pager-266720.aero_modeled.birds_eye_Beam_DF`
    )
group by overall_status
having percent > 1.5

union all

select 'AERO' as dataset, overall_status, count(overall_status) as count, ((count(overall_status) / 5098) * 100) as percent
from `probable-pager-266720.nih_modeled.clinical_studies_main_Beam_DF` csm
inner join `probable-pager-266720.nih_modeled.interventions_Beam_DF` int
on csm.nct_number = int.nct_number
where (lead_sponsor_agency_class = 'Industry') and
    (int.intervention_type = 'Drug') and
    lead_sponsor_agency in (
    select distinct sponsor
    from `probable-pager-266720.aero_modeled.birds_eye_Beam_DF`
    )
group by overall_status
having percent > 2
order by dataset, percent desc
```



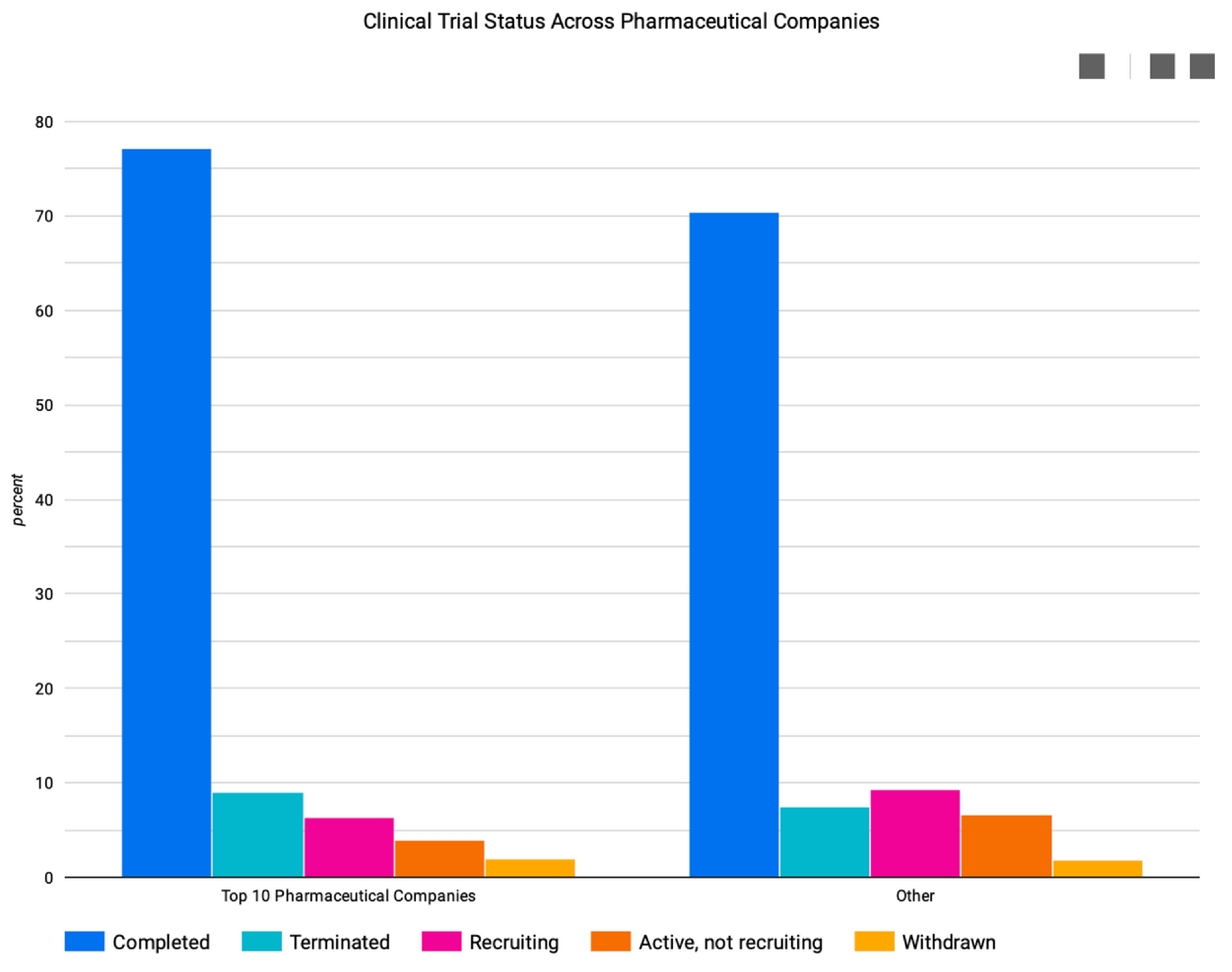
```
%%bigquery

select 'CSM' as dataset, condition, count(condition) as count, ((count(condition) / 38396) * 100) as percent
from `probable-pager-266720.nih_modeled.clinical_studies_main_Beam_DF` csm
inner join `probable-pager-266720.nih_modeled.interventions_Beam_DF` int
on csm.nct_number = int.nct_number
where (lead_sponsor_agency_class = 'Industry') and
    (int.intervention_type = 'Drug') and
    condition in (
    select condition
    from `probable-pager-266720.nih_modeled.clinical_studies_main_Beam_DF` csm
    inner join `probable-pager-266720.nih_modeled.interventions_Beam_DF` int
    on csm.nct_number = int.nct_number
    where (lead_sponsor_agency_class = 'Industry') and
        (int.intervention_type = 'Drug') and
        (condition != 'Healthy') and
        (condition != 'Healthy Volunteers')
    group by condition
    order by count(condition) desc
    limit 10
    ) and
    lead_sponsor_agency not in (
    select distinct sponsor
    from `probable-pager-266720.aero_modeled.birds_eye_Beam_DF`
    )
group by condition

union all

select 'AERO' as dataset, condition, count(condition) as count, ((count(condition) / 4787) * 100) as percent
from `probable-pager-266720.nih_modeled.clinical_studies_main_Beam_DF` csm
inner join `probable-pager-266720.nih_modeled.interventions_Beam_DF` int
on csm.nct_number = int.nct_number
where (lead_sponsor_agency_class = 'Industry') and
    (int.intervention_type = 'Drug') and
    condition in (
    select condition
    from `probable-pager-266720.nih_modeled.clinical_studies_main_Beam_DF` csm
    inner join `probable-pager-266720.nih_modeled.interventions_Beam_DF` int
    on csm.nct_number = int.nct_number
    where (lead_sponsor_agency_class = 'Industry') and
        (int.intervention_type = 'Drug') and
        (condition != 'Healthy') and
        (condition != 'Healthy Volunteers')
    group by condition
    order by count(condition) desc
    limit 10
    ) and
    lead_sponsor_agency in (
    select distinct sponsor
    from `probable-pager-266720.aero_modeled.birds_eye_Beam_DF`
    )
group by condition
```

Airflow DAGs

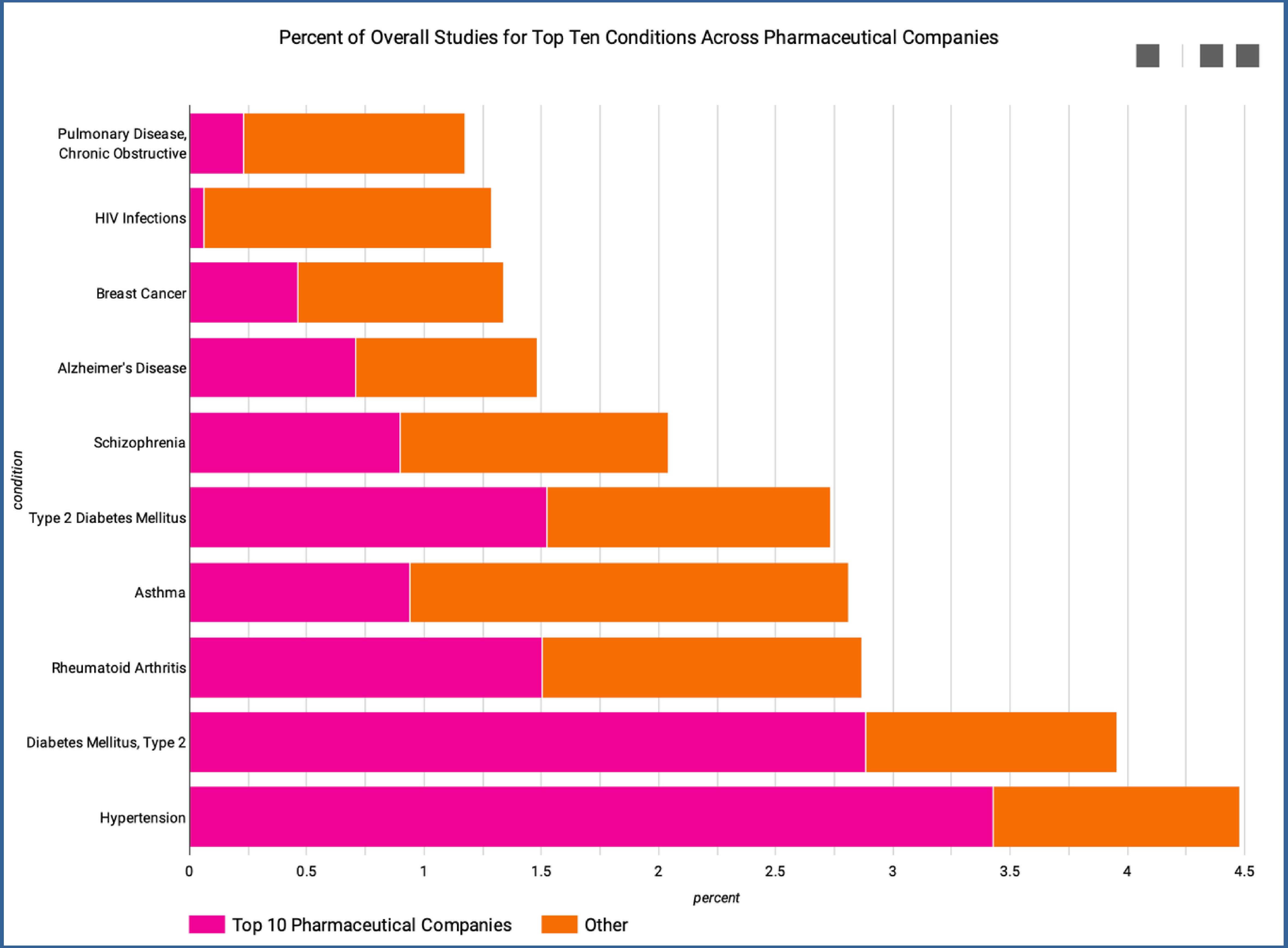
DAGs: Create Dataset DAGs / Load Tables DAGs / Dummy Operators DAGs
Create Tables DAGs / Dataflow DAGs



```
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
```



Clinical Trial Status Across Pharmaceutical Companies
percent
80
70
60
50
40
30
20
10
0
Top 10 Pharmaceutical Companies
Other
Completed
Terminated
Recruiting
Active, not recruiting
Withdrawn



Percent of Overall Studies for Top Ten Conditions Across Pharmaceutical Companies

Conclusions

01 The top ten pharmaceutical companies have slightly more completed clinical trials than other companies, but are similar across other status categories.

02 Although the top ten pharmaceutical companies do oversaturate some drug markets, it is not a significant amount across the top ten conditions.

Future Improvements

01 Group together similar medical conditions.

02 Compare drug trials between sponsor agency classes. (e.g., NIH, US Fed, Other)

Questions?