Washington State Flu Pipeline Tables and Schema

Team Members: Thomas England, Andrew Fuerst, Addison DeSalvo

This database schema follows a star schema design consisting of two reference tables and three fact tables, integrating data from RHINO, FluView, and Census sources to support analysis of respiratory illness trends in Washington State. The reference tables establish the geographic and temporal foundations for the other three tables. (1) The County and Region Reference Table unifies county names, ACH healthcare regions, and population density data from the Census to provide a consistent geographic lookup. (2) The Temporal Reference Table aligns RHINO reporting weeks with CDC epidemiological week data to create a standardized weekly time reference spanning from the 2023 flu season to the present.

The three fact tables draw on these reference dimensions to support analytic comparison across place and time. (3) The County Weekly Illness Comparison Table links RHINO and FluView data to compare county-level illness rates to statewide influenza-like illness percentages. (4) The Healthcare Utilization and Community Burden Table aggregates hospitalization and emergency visit data from RHINO and combines it with population density metrics to evaluate healthcare usage across counties. (5) The Historical Flu Season Summary Table uses FluView and Census data to examine long-term flu trends, identifying annual peaks in flu activity and relating them to changing population densities across decades.

All tables are normalized to Third Normal Form to preserve data integrity, eliminate redundancy, and ensure consistent relationships between temporal, geographic, and epidemiological information. The corresponding entity-relationship diagram will be developed in Draw.io after team review, correction and final approval of the schema.

Format for each table is as follows:

- Datasets Used: the datasets used to create the table
- Purpose: What potential uses this table could have in analytical or predictive models
- Primary Key: The primary key that uniquely identifies each other aspect of the table
- Foreign Key(s): Any key(s) the table shares with other tables that would allow them to possibly be linked
- Third Normal Form Check: Logic for checking if the table is in third normal form
- Notes: Anything of note about the table.
- Highlighted Rows: Any row highlighted in yellow is a column that must be created from an combination of other columns from the datasets. Logic/methodology for creating those columns will be in the Column Source section of each table description
- Table Description: A description of each column for a table, its source dataset and a description will be provided in the tables below so table creation is mapped out. The data type of the column is listed in parentheses next to the column name

Table 1: County and Region Reference Table

- Datasets Used: RHINO + Census
- Purpose: Gives a unified geographic lookup joining county names to population density and ACH locations. It acts as a reference table for all county or geographic based information.
- Primary Key: county_id (county_name if choosing not to make county_id)
- Foreign Key(s): None
- Third Normal Form Check: All non-key attributes depend directly on county_id. No transitive dependencies (ach_region and population_density_2020 both describe the county, not each other).

Column (Data Type)	Column Source	Column Description
county_id (integer)	Created: Not a needed column but is useful as a primary key in other tables. This column is created by linking an integer id number to each county name in the census table and then linking those same id numbers to the counties listed in the RHINO data (determined from ACH codes). This number is used as a reference to counties in other tables.	Created from census data to make a unified numeric reference for all counties. Surrogate key to uniquely identify each county. Not strictly required if county_name values are fully standardized with county (Census and RHINO columns) but retained for best practice and scalability.
county_name (text)	RHINO (county column) and Census (County Name column)	Cleaned and standardized county names (so the county column from both data sets match) from both the RHINO data (this column is present in the github code as a created table linking ACH code with care providers to determine the counties they are in) and county names from the census data. Primary Join Key.
ach_region (text)	RHINO (Location)	Accountable Community of Health. This is where the healthcare providers are listed in the RHINO data. Have as a single column with commas separating each provider if a county has more than one,

		matched to proper county. Counties not mapped (ie. Counties that are in the census data but not found in the RHINO data) to an ACH in the RHINO should be assigned NULL or 'Unassigned' values
population_density_2020 (float)	Census (population Density 2020)	Density of people per county via the 2020 data
JOIN Columns:	RHINO: county	Census: county_name
JOIN Logic:	County names from the RHINO dataset (derived from ACH region mapping) are to be standardized to match the Census County Name field. Each RHINO county entry corresponds to a single Census county name, allowing population density data to be joined directly to RHINO's geographic information. Counties not represented in the RHINO data should be assigned a value of NULL or "Unassigned" to preserve data integrity.	

Table 2: Temporal Reference

- Datasets Used: RHINO + FluView
- Purpose: Aligns CDC epiweeks with RHINO week start and end dates to establish a unified temporal reference for all time-based joins.
- Primary Key: epiweek_id
- Foreign Key(s): None
- Third Normal Form Check: Every non-key field depends only on week_id. There are no derived or repeating fields. There are no dependency between non-key attributes (e.g., season doesn't determine week_start or vice versa).
- Note: The columns epiweek_id and epiweek will match but the idea is to turn the RHINO dates (using the week_start and week_end columns) into epiweek format meaning a six digit integer where the first 4 digits are the year and the last two are the week number. Once done the epiweek column in here is simply to ensure that all time data from both FluView and RHINO data are in this table and it can be removed or ignored and is overall redundant. Just make sure that the FluView epiweek and epiweek_id match one another perfectly so each piece of data is dated and matched properly in future tables.

Column (Data Type)	Column Source	Column Description
epiweek_id (integer)	Created: A surrogate key derived from the RHINO dataset week data (week start/week end) and is to be matched with CDC epiweek data. Epiweek is listed as a six digit number where the first 4 digits are the year and the last two	Unique date identifier in RHINO data and CDC data from RHINO week start and week end and CDC epiweek. Epiweek is coded from six digits, the first 4 being the year and the last two the

	the week id of the year. The goal of this column is to match the weeks listed as week start/week end in the RHINO data and transform them into the same format as epiweek so the two can be easily joined into a single column in future tables. So if the week start/week end in RHINO is 01/01/2025 – 01/07/2025 its epiweek number would be 202501 while the last week of the year 2025 would be 202552 (assuming standard 52 week year)	week of the year. These can be used to match epiweek to week start/end in RHINO data.
week_start (date)	RHINO	The start of a week in a given year to be connected to a specific epiweek_id
week_end (date)	RHINO	The end of a week in a given year to be connected to a specific epiweek_id
cdc_epiweek (integer)	FluView: The epiweek column in the FluView data, it should match the epiweek_id and is here as a check to make sure the two are equal. This column is in a sense a duplicate of epiweek_id and can be ignored or removed so long as the RHINO week_start and week_end dates are in epiweek format.	CDC epidemiological week number, to be matched with epiweek_id. This column is redundant and can be removed if decided as the epiweek format is already a clean date format and it makes sense to transform the RHINO dates into the epiweek format over making a whole new id system.
Season (text)	RHINO: The actual season years the given case is part of (20XX-20YY)	Flu or illness surveillance season provided via RHINO data for a given week
JOIN Columns:	RHINO: week_id	FluView: epiweek
JOIN Logic:	Each CDC epiweek value is to correspond to a RHINO date entry for that same start date range to give a single epiweek_id used for both RHINO and FluView dates. FluView data from the API is on a week-by-week basis allowing for proper correlation	

Table 3: County Weekly Illness Comparison

- Datasets Used: RHINO + FluView (Tables 1 and 2 join through)
- Purpose: Shows weekly county-level illness rates from RHINO compared to statewide FluView ILI (Influenza-Like Illness) activity. Enables comparisons between local and statewide respiratory illness trends over time.

- Primary Key: Composite Key (epiweek_id, county_id, respiratory_illness_type, care_type)
- Foreign Key(s) epiweek_id, county_id,
- Third Normal Form Check: Every attribute depends on the full composite key. No partial dependencies (since county_percent, care_type, etc. only make sense for that full combination).
- Note: This table is joined through Tables 1 and 2 using county_id and epiweek_id from those reference tables. Each week has one row per county. All rows for a given week share the same state_ili_percent value from FluView, while county-level metrics differ by county_id, respiratory_illness_type, and care_type.

Column (Data Type)	Column Source	Column Description
epiweek_id (integer)	Table 2	FK for time reference from table 2
county_id (integer)	Table 1: one row per county within each epiweek period (same format as the RHINO dataset)	FK for geographic reference from table 1
respiratory_illness_type (text)	RHINO (Respretory Illness column)	Flu, COVID-19, RSV. Only if we choose to go outside of just flu but it could be useful to have all three for comparison of what is flu and what is not
care_type (text)	RHINO (Care Type column)	Whether case was hospitalization or an emergency visit
county_ili_percent (float)	RHINO (1-Week Percent column)	Weekly illness percentage per county, illness, and care type.
state_ili_percent (float)	FluView (wili column)	Weighted ILI (Influenza like illness) percentage for the state that week
deviation_from_state_average (float)	Created: Difference between state and county averages (state_ili_percent – county_ili_percent)	state_ili_percent – county_ili_percent showing how county values differ from the statewide average.
JOIN Columns:	RHINO: epiweek_id - formulated from week_start and week_end in RHINO data, derived through Table 2 (Temporal Reference) and matched with CDC epiweek	FluView: epiweek
JOIN Logic:	This table compares weekly county-level illness rates from RHINO with statewide ILI percentages from FluView. Each record represents a unique combination of county_id and week_id, joined THROUGH Tables 1 (County Reference) and 2 (Temporal Reference). All rows for a given week_id share the same	

	state_ili_percent value from FluView, while county-specific metrics come from RHINO. The result supports county-by-county comparison of local versus statewide illness trends for each week.

Table 4: Healthcare Utilization and Community Burden

- Datasets: RHINO + Census (Table 1 join through)
- Purpose: Compares hospitalization and ER visit percentages for respiratory illnesses, adjusted by county population density, using the RHINO dataset as a whole from the 2023-2024 flu season through the current 2025-2026 flu season to look at resource usage and community burden.
- Primary Key: county_id
- Foreign Key(s): county_id
- Third Normal Form Check: Each attribute depends solely on county_id.. There are no transitive or partial dependencies.. The derived field (hospital_to_er_ratio) is expected in a summary table and does not break 3NF.
- Note: Each row represents a single county-level aggregate (one summarizing row per county), summarizing hospitalization and emergency visit activity across all weeks in the RHINO dataset.

Column (Data Type)	Column Source	Column Description
county_id (integer)	Table 1	FK for county and ACH locations
population_density_2020 (float)	Census (2020 Population Density column)	Population density of a given county.
hospitalization_percent (float)	Created: Made from a mix of the RHINO 1-week percent + care type = hospital to determine the percentage that end up hospitalized within each county. Make sure each entry is county specific and covering the entire timeline of the RHINO dataset.	The percent of people in the county hospitalized by taking all values from RHINO column Care Type which is equal to hospitalization. Leave empty if no matching cases for a given county from census data (for counties that no ACH was listed for in the RHINO data)
er_visit_percent (float)	Created: Made from a mix of the RHINO 1-week percent + care type = er visit to determine the percentage that end up hospitalized within each county. Make sure each entry is county specific and covering the entire timeline of the RHINO dataset.	The percent of people in the county who visit the ER for a respiratory illness using values where care type is equal to Emergency Visits. Leave empty if no matching cases for a given county from census data (for counties that no ACH was listed for in the RHINO data)

hospital_to_er_ratio (float)	Created: Simply a ratio for each county that is calculated as hospital_percent / er_visit_percent from each county to determine resource utilization (in terms of how many are hospitalized vs simple er visits)	The ratio of hospitalizations to er visits for respiratory like illnesses. This column will show whether ER or hospital usage dominates in certain counties. Leave empty if no matching cases for a given county (again for counties in Census data that did not match an ACH value in the RHINO data).
JOIN Columns:	RHINO: County	Census: County Name
JOIN Logic:	RHINO county-level aggregates are joined with Census population density via Table 1's unified county references. Each county appears once, showing cumulative healthcare utilization metrics alongside its 2020 population density. This enables comparison of how population relates to hospitalization and ER visit rates across Washington State counties and which counties have the highest or lowest ER/Hospitalization ratios.	

Table 5: Historical Flu Season Summary

- Datasets: FluView + Census
- Purpose: Provides a historical look of flu seasons using FluView weekly data joined with Census population data to see how flu cases change based on population. FluView goes back to around 1997 if more data wants to be used on a more week by week case (historic pull) or it can be consolidated into a more yearly view
- Primary Key: year
- Foreign Key(s): None (no shared values with other tables)
- Third Normal Form Check: All fields depend on the PK (year). There are no transitive or partial dependencies. peak_vs_avg_diff is a derived analytic field and acceptable denormalization for performance, not structural redundancy.
- Note: Data can be drawn to look at yearly averages for flu using epiweek and 52 week intervals in which all the data for that year is averaged out and combined with the population data of that decade. RHINO data cannot be used here as it only goes back to 2023. Each row should represent one year of time.

Column (Data Type)	Column Source	Column Description
Year (integer)	Created: Derived from FluView through epiweek or from epiweek_id in Table 2 (though I recommend using	The given year the FluView data is taken from using epiweek. The six digit code is a combination of year and

	the actual FluView epiweek as it goes back further than our temporal table will which only goes back to 2023). Use the first four digits of the epiweek number to get the year for the year column	week where the first four digits are the year and the last two is the week. Use the first four digits of epiweek to create a year column
decade_year (integer)	Created: Taken from Census data. For example, if the decade year is for times between 2000 and 2010 then decade year should be 2000.	The decade census data matched so that the year column is within the decade range defined
population_density_decennial (float)	Census: Population Density	Population density for that decade to be used as population standard of the state. Use total population density/number of residents of the state to see how rate of illness changes with population at least on a small scale (1990, 2000, 2010, 2020)
peak_week_id (integer)	Created: The week number of that year with highest WILI value. That is to say that this will be the value from FluView in that given year (all epiweeks sharing the same first 4 digits that define year) that defines the week with the highest Wili value for that year. Use the last two numbers of the epiweek value for that week to define the week.	Use highest WILI (weighted influenza like illness) to define peak week of that year. Gives the week that had the highest confirmed cases for that year.
peak_ili_percent (float)	Created: Simply the value from the Wili column of the peak week.	The WILI number for the peak week to see what the highest weighted influenza like illnesses for that year was
average_wili_percent (float)	Created: From FluView, this is the averaged WILI number for the given year averaged across all epiweeks of that year. That is to say the averaged (mean) value of all epiweeks sharing the same	The average Weighted Influenza Like Illnesses for the given year, averaged out from all the WILI values for all epiweeks of a given year. This is simply the mean of all the WILI of the given year

	first 4 digits in the epiweek code to get an average wili value for that year	(sum of WILI/number of weeks)
peak_vs_avg_diff (float)	Created: The between the peak value of the year and the average value of the year. Simply created by calculating the difference between peak_ili_percent and average_wili_percent for each year. (peak_ili_percent – average_wili_percent)	The difference (peak-average columns) between what the peak WILI value for the year was compared to the average for a proper comparison of how large a difference actual "flu season" is from the rest of the year.
JOIN Columns:	FluView: epiweek	Census: year (decade)
JOIN Logic:	FluView's annual summaries (created columns giving the average and peak flu intensity) are joined to the corresponding Census decade's population density via the decade reference. Each row represents a single flu year, showing average and peak illness intensity relative to the decade's population baseline.	