**Whitepaper: Impact of an influenza pandemic on agricultural workers and food production in the US**

**1. Executive Summary**

This document describes a compartmental epidemiological modeling framework designed to simulate disease transmission within a household-structured population. The model distinguishes between two subpopulations, a "community" population and an "agricultural worker" population, to capture distinct demographic and social mixing patterns. The framework leverages publicly available data from the U.S. Census Bureau's American Community Survey (ACS) and the National Agricultural Workers Survey (NAWS) to parameterize the model at a granular, county-level resolution. The analysis is orchestrated by a series of R scripts, which handle data importation, model calibration, simulation, and output visualization.

**2. Introduction**

Understanding the dynamics of disease spread requires models that can account for the heterogeneity of human populations. This framework focuses on two key aspects of heterogeneity: household structure and occupational demographics. By explicitly modeling these factors, the framework aims to provide a more nuanced perspective on disease transmission, particularly in the context of agricultural communities which may have different household sizes, crowding levels, and contact patterns compared to the general population. The ultimate goal is to assess the potential impact of a disease outbreak on both human health and related economic sectors like food production.

The provided epidemiological model is a direct extension of the **House & Keeling model**, building upon its core concepts in three key ways:

### 2.1. Foundational Household Structure

Like the original House & Keeling model, this model uses a **household-structured approach** to account for different transmission dynamics within and between households. It distinguishes between a high, within-household transmission rate (1) and a lower, between-household transmission rate (2).3 This core principle recognizes that close, frequent contact within a household leads to a higher probability of infection compared to more random, less frequent contact in the wider community.4

### 2.2. Incorporation of Demographic Heterogeneity

The model builds upon the foundational concept by introducing **demographic heterogeneity**, specifically through the inclusion of a two-population structure: a general community and agricultural workers. This is a significant expansion from the original model, which typically considers a single, homogeneous population. By separating the population into two distinct groups, the model can account for different household sizes, living conditions, and mixing patterns between the two populations, a crucial step for accurately modeling disease spread in a non-homogeneous setting.

### 2.3. Advanced Parameterization with Real-World Data

This model takes the theoretical framework a step further by using **real-world data** to parameterize its components. It uses data from the American Community Survey (ACS) to define the distribution of household sizes and the National Agricultural Workers Survey (NAWS) to identify the characteristics of agricultural worker households. Crucially, it introduces a **crowding factor** () to modify the within-household transmission rate. This factor is derived from the NAWS data and is applied to households identified as "crowded," directly linking the model's transmission parameters to empirical observations of living conditions and thereby improving its predictive capability.

In summary, the model uses the House & Keeling framework as its foundation, then extends it by incorporating a two-population structure and grounding its parameters in detailed, real-world demographic and occupational data.

**3. Modeling Framework**

The core of the framework is an epidemiological model defined using the odin package in R. The model, named household\_model\_twopop\_crowding, is a two-population extension of a standard household SIR (Susceptible-Infectious-Recovered) model.

* **Compartments**: The model tracks the number of households in each possible state (defined by the number of susceptible, infected, and recovered individuals) for each of the two populations: **Community (C)** and **Agricultural Workers (A)**.
* **Disease Dynamics**: The model incorporates two primary modes of transmission:
  + **Household Transmission (tau)**: Transmission within households, with a basic attack rate that can be adjusted for crowded households (tau\_boost).
  + **Community Transmission (beta)**: Transmission between households, representing contact in the wider community.
* **Mixing**: The model allows for a flexible mixing parameter (eps) between the two populations. An eps of 0 indicates perfectly assortative mixing (populations only interact with themselves), while an eps of 1 indicates proportional mixing (populations mix freely). The run\_analysis.R script sets eps = 0, indicating that the populations are treated as separate communities with no cross-interaction.
* **Model Implementation**: The model is defined in the epimodels.R script, which also includes simpler models for comparison (basic\_model, household\_model).

**4. Data and Parameterization**

The model is parameterized using a combination of national, regional, and county-level data.

* **U.S. Census Bureau ACS Data (import\_acs.R)**: This script downloads and processes county-level data on household sizes and crowding. It uses the tidycensus package to access the Census API, requiring a valid API key. It also incorporates urban/rural data to further characterize each county.
* **NAWS Data (import\_naws.R)**: This script reads data from a SAS file (naws\_all.sas7bdat) to derive regional-level distributions of household sizes and crowding specifically for agricultural workers.
* **Model Calibration (calibrate\_model.R)**: The calibrate\_model.R script uses the national summaries of ACS and NAWS data to calibrate the household\_model\_twopop\_crowding model before it is run at the county level. This step establishes baseline parameter values.

**5. Simulation and Analysis**

The simulation and analysis workflow is automated by the run\_analysis.R script.

* **Main Workflow (run\_analysis.R)**: This script acts as the master control, sourcing other scripts in the correct order: calibrate\_model.R, simulate.R, and finally plot\_model\_output.R.
* **County-level Simulation (simulate.R)**: This script iterates through each U.S. county. For each county, it retrieves the appropriate data on household size and crowding distributions and then runs the household\_model\_twopop\_crowding to simulate disease spread within that county's specific demographic profile. The results are saved to epidf\_indiv\_full.csv.
* **Visualization (plot\_model\_output.R)**: The plot\_model\_output.R script reads the simulation results from the CSV file and generates a variety of plots and maps. These visualizations include time-series plots of infection curves for each subpopulation and geographic maps showing spatial patterns in disease outcomes across counties.

**6. Development Notes and Discussion**

The project notes from the README.md file offer critical insights into the development process and future directions.

* **Modeling Decisions**: The notes indicate a strategic shift away from a full spatial transmission model. The rationale is that the timing and location of a future pandemic's introduction are highly uncertain, making a comprehensive spatial model less useful than a more focused county-level analysis of local impact.
* **Data Imputation**: The notes raise a key challenge: the lack of county-level data on agricultural worker demographics. While the current approach uses regional NAWS data to inform county-level models, a more sophisticated approach, such as a pooled Bayesian model, is suggested as a potential improvement to impute these values more credibly.
* **Future Work**:
  + **Outcome Metrics**: The notes highlight the need for a clearer definition of outcome metrics beyond just disease spread, suggesting a link to downstream impacts on food production and availability, possibly by integrating with the "FoodTwin" project.
  + **Parameterization**: The notes also mention the need to validate the beta parameterization and to explore the variability of R0 across locations using the current methodology.
  + **Crowding**: The notes discuss the complexities of implementing household crowding, particularly for agricultural workers. The current approach assumes a certain relationship between general crowding and crowding among agricultural workers, but alternative strategies are proposed, such as assuming the same standard deviation in crowding rather than the same relative difference.

**7. Conclusion**

This epidemiological modeling framework represents a robust tool for simulating disease spread within a heterogeneous, household-structured population. By combining rich demographic data with a flexible modeling approach, it provides a powerful means to explore the potential impacts of a pandemic at a local level. The development notes serve as a valuable roadmap for future refinements, particularly in improving data parameterization and expanding the scope of the analysis to include broader economic impacts.