# **Problem description**

Your goal is to predict how likely individuals are to receive their H1N1 and seasonal flu vaccines. Specifically, you'll be predicting two probabilities: one for h1n1\_vaccine and one for seasonal\_vaccine.

Each row in the dataset represents one person who responded to the National 2009 H1N1 Flu Survey.

**Labels**

For this competition, there are two target variables:

* h1n1\_vaccine - Whether respondent received H1N1 flu vaccine.
* seasonal\_vaccine - Whether respondent received seasonal flu vaccine.

Both are binary variables: 0 = No; 1 = Yes. Some respondents didn't get either vaccine, others got only one, and some got both. This is formulated as a multilabel (and *not* multiclass) problem.

**The features in this dataset**

You are provided a dataset with 36 columns. The first column respondent\_id is a unique and random identifier. The remaining 35 features are described below.

For all binary variables: 0 = No; 1 = Yes.

* h1n1\_concern - Level of concern about the H1N1 flu.
  + 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.
* h1n1\_knowledge - Level of knowledge about H1N1 flu.
  + 0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.

**Behavioural Data**

* behavioral\_antiviral\_meds - Has taken antiviral medications. (binary)
* behavioral\_avoidance - Has avoided close contact with others with flu-like symptoms. (binary)
* behavioral\_face\_mask - Has bought a face mask. (binary)
* behavioral\_wash\_hands - Has frequently washed hands or used hand sanitizer. (binary)
* behavioral\_large\_gatherings - Has reduced time at large gatherings. (binary)
* behavioral\_outside\_home - Has reduced contact with people outside of own household. (binary)
* behavioral\_touch\_face - Has avoided touching eyes, nose, or mouth. (binary)

**Doctor Recommendations**

* doctor\_recc\_h1n1 - H1N1 flu vaccine was recommended by doctor. (binary)
* doctor\_recc\_seasonal - Seasonal flu vaccine was recommended by doctor. (binary)
* chronic\_med\_condition - Has any of the following chronic medical conditions: asthma or an other lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. (binary)

**Miscellaneous Data**

* child\_under\_6\_months - Has regular close contact with a child under the age of six months. (binary)
* health\_worker - Is a healthcare worker. (binary)
* health\_insurance - Has health insurance. (binary)

**Opinion Data**

* opinion\_h1n1\_vacc\_effective - Respondent's opinion about H1N1 vaccine effectiveness.
  + 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
* opinion\_h1n1\_risk - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine.
  + 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
* opinion\_h1n1\_sick\_from\_vacc - Respondent's worry of getting sick from taking H1N1 vaccine.
  + 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
* opinion\_seas\_vacc\_effective - Respondent's opinion about seasonal flu vaccine effectiveness.
  + 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
* opinion\_seas\_risk - Respondent's opinion about risk of getting sick with seasonal flu without vaccine.
  + 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
* opinion\_seas\_sick\_from\_vacc - Respondent's worry of getting sick from taking seasonal flu vaccine.
  + 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.

**Personal Data**

* age\_group - Age group of respondent.
* education - Self-reported education level.
* race - Race of respondent.
* sex - Sex of respondent.
* income\_poverty - Household annual income of respondent with respect to 2008 Census poverty thresholds.
* marital\_status - Marital status of respondent.
* rent\_or\_own - Housing situation of respondent.
* employment\_status - Employment status of respondent.
* hhs\_geo\_region - Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
* census\_msa - Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.

**Household Data**

* household\_adults - Number of *other* adults in household, top-coded to 3.
* household\_children - Number of children in household, top-coded to 3.

**Employment Data**

* employment\_industry - Type of industry respondent is employed in. Values are represented as short random character strings.
* employment\_occupation - Type of occupation of respondent. Values are represented as short random character strings.