# Methods Section

Thomas Hsiao

## Methods

Once the sessions were filtered by the aforementioned inclusion criteria, we subsetted the data further to only sessions with a response to the self-report high blood pressure diagnosis question ($n = 1,280,717$). To de-duplicate and prevent overcounting of the same individual, we first split all sessions into two groups: 1) those associated with an account ID ($n = 417,291$) and 2) those without ($n = 863,426$). The Pursuant kiosk allows users the option to create an account, which can be used to track responses and measurements over time upon login before each session. We assume account ID is an unique identifier for individuals. Since county and single-year are the spatiotemporal unit of analysis for our estimates of hypertension prevalence, we make an implicit *de facto* assumption and collapsed all sessions under a single account ID-county-year combination into one observation. The hypertension indicator was set to true if over half of all sessions under the account ID-county-year were deemed hypertensive.

For the second group of sessions with no associated account ID, we relied on the pseudo-ID to approximately identify individuals. The pseudo ID is a unique identifier indexed by DOB, ethnicity, gender, and kiosk location. Within a year at a specific kiosk location, the chances of two separate individuals of the same gender and ethnicity sharing a birthday are reasonably low (consider the famous birthday problem while also including two genders and five racial ethnic groups). We then proceeded with deduplication in the exact same way as that for the sessions with account IDs described in the previous paragraph, but replacing account ID with pseudo ID. While the account ID data revealed the possibility of an individual recording data at multiple kiosk locations (which the pseudo ID would count as multiple different observations), we already made the *de facto* assumption to count individuals for where they are, not to tie them to a single location. In addition, account holders appearing in multiple locations was only the case for $2,127$ out of $385,627$ ($\sim 0.005\%$) of all unique account ID-county-years and we believe ignoring this overcounting would have a small effect on the analysis.

Following deduplication, the final dataset used for analysis consisted of $n = 1,210,339$ observations indexed by individual-location-years.

We conducted a sensitivity analysis to evaluate robustness of the deduplication. Instead of aggregating over county among sessions with account ID, we aggregate only over account ID-year and set the location of the collapsed observation to the county with the greatest number of measurements for the account in the given year. For ties between counties, we randomly choose one county to represent the observation with equal probability. The hypertension indicator is then defined. The assumption in this case is *de jure* where we assume each individual is only associated with their location of "residence" determined by which location they frequent the most. We conducted another sensitivity analysis of the sessions with no account ID by not performing any aggregation and assuming each session denotes a unique individual. While more sophisticated algorithms to detect repeated measurements exist (such as factoring in time at kiosk use), the modeling choice in the main analysis and the sensitivity analyses here represent two extremes and should capture the range of estimates possible under the possible assumptions to be made. The number of total sessions with multiple associated IDs (n=XX) also only compose XX% of all sessions, and their effects are not likely to be felt at the national or even state level.

## Results