

How We Ensure Data Privacy

We follow national and state guidelines to assure the data safety of Kaiser Permanente Southern California's (KPSC) Electronic Health Records. We follow a multi-step process for data de-identification to safeguard our data against any accidental disclosure.

To begin, we follow the guidelines of the Department for Health and Human Services (DHHS) for de-identification.¹ First, we exclude data from KPSC members who are 85 years or older, as this population group tends to be small. We then use high-level aggregation of individual level data into large population groups by choosing a large geographic unit for our map (Health Districts) and broad age groups (18-44, 45-64, 65-85), so that no single individual can be identified from our data. This is also the reason we only present the prevalence for the biggest race/ethnic groups in LA County (Hispanic, White, Asian or African American/Black) and do not present the prevalence for smaller groups or combinations.

Next, we assess the size of the denominator and numerator that each prevalence estimate is based on and its potential risk for inadvertent disclosure of any individual's hypertension diagnosis. For this purpose, we followed federal guidelines and conducted a statistical expert review and made determinations of potential disclosure risk.² To add additional safeguards, our statistical expert applied a data protection method called a *publication score* to identify hypertension prevalence estimates that are based on population denominators or numerators that are considered too small to be released safely. The publication score is a method proposed by the California Department of Public Health as well as other health departments, that assesses and mitigates potential risk of disclosure. A publication score is calculated for each health district-year-age-sex-race/ethnicity combination. In our data, publication scores range from -15 to 41, with higher positive risk scores indicating a higher potential risk of disclosure.³ Any prevalence estimate with a publication score of 12 or less is considered a minimal risk of potential disclosure per California State Health Department guidelines.

Specifically, the publication score considers the size of the underlying population of a particular age-sex-race/ethnicity combination within a Health District, the number and size of age groups, the number of race/ethnic categories and whether prevalence is combined across multiple years or presented for single years. For example, presenting prevalence for 1 year age groups would add 7 points to the publication score, whereas using broad age groupings of 10+ years (like the ones used here) only adds 2 points to the score. For prevalence estimates that are, for example, not broken down by age at all, no points are added to the score. We follow State Department of Health Guidelines and suppress all prevalence estimates that have a publication score of 12 or higher.⁴ The final dataset only includes the name of the Health District, the age-sex-race/ethnicity-year category, and the aggregated hypertension prevalence estimate for each respective sub-population.

¹ <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

² <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

³ <https://www.dhcs.ca.gov/dataandstats/Documents/DHCS-DDG-V2.0-120116.pdf>, p.27

⁴ <https://www.dhcs.ca.gov/dataandstats/Documents/DHCS-DDG-V2.0-120116.pdf>, p.17

No identifiable individual information, and no information on the actual numerators or denominators for any hypertension prevalence estimate are included in our database.