A tool to find the optimal penetration rate of a health system to estimate the disease prevalence

User manual

Piaopiao Li

Advisor: Dr. Hui Shao

Objective

There is an issue of overestimating the prevalence of disease using EHR (Electronic Health Records). In order to solve the problem, a health system penetration rates (HSPR)-based method was proposed to estimate the prevalence of disease. The observed health system penetration rates were estimated as a ratio of total patients who visited the healthcare systems (e.g., OneFlorida network) to the total number of residents in each zip code area. The HSPR-based method assumed that: the disease prevalence in a zip code area measured by EHR will increase as the penetration rate of this area gets higher. When the HSPR is higher than a cutoff threshold, the trend line between prevalence and penetration rate will be plateaued. We can then produce accurate prevalence estimation using data from areas where the observed HSPR is higher than the cutoff threshold.

The optimal penetration rate of a health system that can be used for estimating the disease may vary among the large health systems located in different geographic areas (e.g., OneFlorida in Florida v.s. University of Pittsburgh Medical Center (UPMC) located in Pennsylvania). The program intends to find the optimal cutoff value of penetration rate where the trend line between prevalence and penetration rate plateaus for different large health systems. The researcher can further use data from areas where penetration rate is higher than the cutoff value to estimate the disease prevalence in that health system. This method can solve the issue of overestimating the prevalence of disease using EHR.

The program is implanted in R software. **Input** of the program includes 1) a dataset contains information about the health system penetration rate of a specific area (e.g., zip code area) and its corresponding disease prevalence measured by EHR. **Output** of the program includes 1) the optimal cutoff value of penetration rate where the trend line between prevalence and penetration rate plateaus, 2) the regression coefficient of the plateaued trendline, 3) the graphs showing the trend.

User directions:

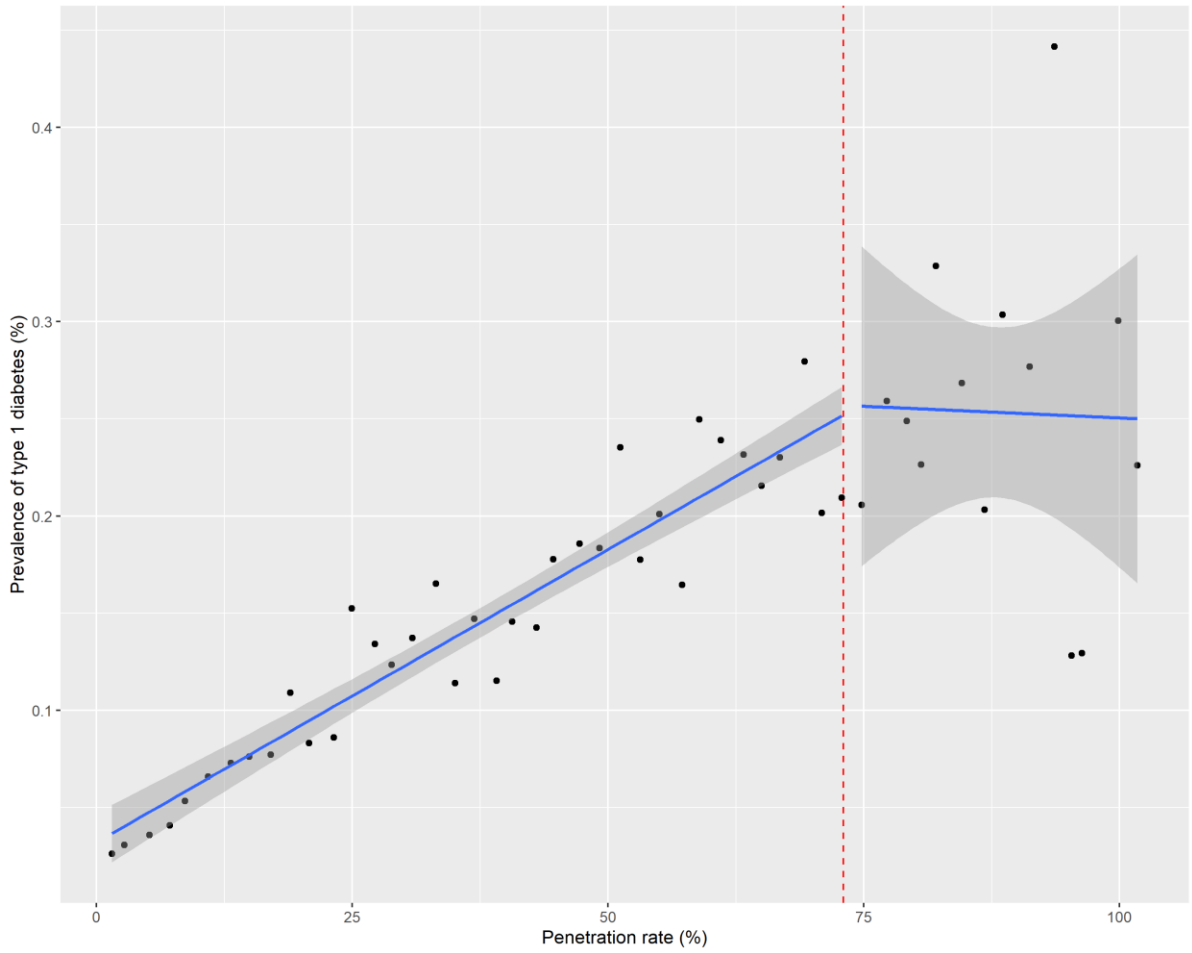
1. Move the R document “Estimate optimal penetration rate” to the file where you store your data.
2. Open the R document.
3. Load your dataset.
4. Click “run” and then get the optimal penetration rate and corresponding plot.

Example:

We will use the data containing the information of OneFlorida health system penetration rate and T1D prevalence measured in the OneFlorida EHR as an example to illustrate how the tool works.

1. Import the attached document “Estimate optimal penetration rate” into your R software.
2. Call the required package “ggplot2”
3. Run the author-defined function “penetration_cutpoint”
4. Load your dataset (you can use sample dataset: “2019_T1D.csv” to have a try)
5. Use the defined function “penetration_cutpoint” to get the optimal cutoff value for penetration rate and corresponding β coefficient for the plateaued trendline
6. Run the author-defined function “plot_penetration”
7. Use the defined function “plot_penetration” to get graph showing trendline and vertical cutoff value line

Using the sample dataset, you can find that when the penetration rate of OneFlorida heath system is bigger than 73%, the prevalence of T1D plateaued (β= 0.00024). The visualization of the results is shown in the figure below. The researchers can further use data from areas where penetration rate higher than 73% to estimate the T1D prevalence and conduct T1D surveillance.



I agree that this program can be made available to collaborators and if deemed of interest to others, for public download at the CoDES website.