Design Document:
Algorithm to Calculate Disease Probability

Phillip F. Tellier

April 2022

1 Setup

Let events $S_1, S_2, S_3 \ldots, S_n$ be the event that various symptoms occur in a study participant. Let events $D_1, D_2, D_3 \ldots, D_n$ be the event that various symptoms occur in a study participant. Then $P(S_1)$ is the probability that a participant has the symptom represented by S_1. Similarly:

- $P(S_1 \cap S_2)$ is the probability that a participant has the symptoms represented by S_1 and S_2.
- $P(S_1|D_2)$ is the probability that a participant has the symptom represented by S_1 given that they have the disease represented by D_1
- $P(S_1 \cap S_2^c)$ is the probability that a participant has the symptoms represented by S_1 and does **not** have S_2.

2 Data Collected

For each study entered into the program, the following data is collected for every disease D_x where $S_1, S_2, S_3 \ldots, S_n$ are all symptoms of the disease D_x:

$$\begin{aligned} P(S_i \mid D_x), &\quad \forall i \in [1,n] \\ P((S_i \cap S_j) \mid D_x), &\quad \forall i,j \in [1,n] \\ P((S_i \cap S_j \cap S_k) \mid D_x), &\quad \forall i,j,k \in [1,n] \\ \text{... all the way up to ...} & \\ P((S_{i_1} \cap S_{i_2} \ldots \cap S_{i_n}) \mid D_x), &\quad \forall i_1, i_2, \ldots i_n \in [1,n] \end{aligned} \tag{1}$$

3 Algorithm

We want to calculate $P(D_x \mid S_1^* \cap S_2^* \cap \ldots \cap S_n^*)$ where * for each S_i^* could be chosen to be S_i^c or S_i *e.g.* we might want the algorithm to calculate $P(D_x \mid S_1^c \cap S_2^c \cap S_3 \cap S_4 \cap S_5)$

We know that

$$P(D_x \mid S_1^* \cap S_2^* \cap \ldots \cap S_n^*) = \frac{P(S_1^* \cap S_2^* \cap \ldots \cap S_n^* \mid D_x) \times P(D_x)}{P(S_1^* \cap S_2^* \cap \ldots \cap S_n^*)} \tag{2}$$

So we need to express $P(S_1^* \cap S_2^* \cap \ldots \cap S_n^*)$ in terms of the collected data. To do this do the following: starting at $m = 1$ note that $P\left(S_1^* \cap S_2^* \cap \ldots \cap S_{m-1}^* \cap S_m^* \cap S_{m+1}^* \ldots \cap S_n^*\right)$ will fill into one of two cases

- case 1: S_m^* is $S_m \implies \mathrm{P}\left(S_1^* \cap S_2^* \cap \ldots \cap S_{m-1}^* \cap S_m \cap S_{m+1}^* \ldots \cap S_n^*\right)$
  and we do not need to rewrite our formula
- case 2: S_m^* is $S_m^c \implies \mathrm{P}\left(S_1^* \cap S_2^* \cap \ldots \cap S_{m-1}^* \cap S_m^c \cap S_{m+1}^* \ldots \cap S_n^*\right)$
  and we can rewrite the formula as
  $P\left(S_1^* \cap S_2^* \cap \ldots \cap S_{m-1}^* \cap S_{m+1}^* \ldots \cap S_n^*\right) - P\left(S_1^* \cap S_2^* \cap \ldots \cap S_{m-1}^* \cap S_m \cap S_{m+1}^* \ldots \cap S_n^*\right)$
  notice that the "c" has disappeared from the S_m^c to become S_m

Increment m and repeat this process on all the terms for the case that applies until finishing with $m = n$. At this point the expression is the sum and differences of the data collected since the "c's" have disappeared.