**S2 File.**

**Calculation of *I*(*Rc*)**.

Step 1: Calculate *p*(*df*), a prior probability of the inference diagnosis *df*:

Step 2: Calculate *p*(*df*|*Rc*), a posterior probability of the inference diagnosis *df* with the reason candidate *Rc*.

Step 3: *pd*(*Rc*) is defined as a difference between *p*(*df*|*Rc*) and *p*(*df*) for the diagnosis *df*:

. (2)

If |*Rc*| = 1, calculate *pd*(*Rc*) based on the equation (2) and *I*(*Rc*) (= *pd*(*Rct*)) is obtained. If |*Rc*| > 1, go to the Step 4.

Step 4: If |*Rc*| > 1, calculate element-wise total positive effect *fp* and a total negative effect *fn*.

(4) (5)

Step 5: Calculate a penalty term *f*(*Rc*) for |*Rc*| > 1 defined as follows:



(3)

Step 6: Calculate *I*(*Rc*) for |*Rc*| > 1 defined as follows:

Step 6 can be alternatively written as:



(6)