**Example: assume that a case control study enrolled m cases and n controls, and exposure 1 is the exposure under consideration**

Step 1: Determine a threshold of Gower distance for matching

To find a threshold, Gower distances between all cases and controls are calculated using all exposures. Then, the maximum bipartite graph algorithm is used to find each case's closest control while ensuring each control is matched to no more than one case. Next, a control is selected randomly for each case. A dataset is constructed:

|  |  |  |
| --- | --- | --- |
| Type of match | Case-control pair | Gower distance |
| Closest | Case 1 – control 3 | 0.123 |
| Closest  | Case 2 – control 5 | 0.456 |
| Closest | …… | …… |
| Closest | Case m – control 2 | 0.789 |
| Random | Case 1 – control 1 | 0.321 |
| Random | Case 2 – control 4 | 0.654 |
| Random | … | …… |
| Random | Case m – control 6 | 0.987 |

A logistic regression on whether the control is the closest or the randomly selected one based on the distance is created (i.e., logit(Pclosest) = β0 + β1 × Gower distance). The logistic regression model can be used to predict the distance corresponding to a given probability. For example, the plot below shows the distance at which the predicted probability is at least 50% that the control is the closest vs. a randomly selected one. Such a distance can be used as the threshold of Gower distance for matching.



Step 2: Determine confounders and intermediates for calculating Gower distances for exposure 1

Exposure a

Exposure b

Exposure c

Exposure d

Exposure e

Exposure 1

Illness status

Exposures a, b, and c are considered confounders and are used to calculate Gower distance in step 3, whereas the exposure of interest (exposure 1) and exposures d and e, considered possible intermediates, are not used to calculate Gower distance.

Step 3: Create analytic data set for exposure 1

**Sequentially apply the following steps:**

1. **Keep case-control pairs that have a Gower distance within the threshold**
2. **Ensure each control is matched with only one case. If two cases share the same controls, merge cases and controls into the same stratum. To favor matching as many cases as possible and make the strata more balanced in size, controls are prioritized to be matched with a case from the stratum with the smallest size. If the sizes of strata were tied, priority is given to the closer case.**
3. **Exclude cases without a control matched. Exclusion of cases may bias the estimates and reduce accuracy and precision of estimates. Exclusion of cases should be balanced with the tightness of the threshold. Sensitivity analysis of using a random sample of cases and all controls may be done to test the impact of exclusion of cases on estimates.**

**Select the 20 closest controls for each case based on Gower distance to ensure strata are well populated after subsequent filtering as described in the right box**

A total of m × 20 case-control pairs are retained:

*Case 1 – Control 1*

*Case 1 – Control 3*

*……*

*Case 2 – Control 2*

*Case 2 – Control 5*

*……*

*Case m – Control 4*

*Case m – Control 6*

*……*

**Calculate Gower**

**distance between**

**each case and each**

**control using exposure set as described above**

Gower distances are calculated for m × n case-control pairs:

*Case 1 – Control 1*

*Case 1 – Control 2*

*Case 1 – Control 3*

*……*

*Case 1 – Control n*

*Case 2 – Control 1*

*Case 2 – Control 2*

*Case 2 – Control 3*

*……*

*Case 2 – Control n*

*Case m – Control 1*

*Case m – Control 2*

*Case m – Control 3*

*……*

*Case m – Control n*

Steps 2 and 3 are repeated to create analytic data set for each exposure. Analytic data sets are then individually analyzed using the standard conditional logistic regression model.