**Appendix A: Clustering families using self-organizing map**

Let () be the familial aggregation vector of the *ith* family in the dataset, where *N* is the number of cancer categories included in the analysis, and *M* is the total number of families. The self-organizing map consists of a regular grid of nodes. Each node is associated with an *N*-dimensional codebook vector. Let  () be the codebook vector of the *jth* node on the map. The training algorithm for forming the *familial aggregation space* is given as follows:

**1**: Present an input vector **x***i* for training at random.

**2**: Find the winning node *s* on the map with the vector **m***s* which is closest to **x**i such that

 

**3**: After the winning node *s* is selected, update the weight of every node in the neighbourhood of node *s* by

 

where is the gain term at time *t* () that decreases in time and converges to 0.

**4**: Increase the time stamp *t* and repeat the training process until it converges.

After the training process was completed, each input vector (i.e. family) was mapped to a grid node closest to it on the self-organizing map. A *familial aggregation space* was thus formed. This process corresponded to a projection of the multi-dimensional input vectors onto an orderly two-dimensional space where the proximity of the input vectors was preserved as faithfully as possible. Consequently, familial similarities, in terms of both the types of extracolonic cancers and the strength of the CRC aggregation were explicitly revealed by their locations and neighbourhood relationships on the map. For all families mapped to a node, a familial risk category, based on family history of cancer, was then revealed by retrieving the codebook vector correspond to a node on the self-organizing map.

**Appendix B: Partitioning the self-organizing map using k-means**

The k-means algorithm used for running on the *familial aggregation space* was as follow:

**1**: Select k nodes from the self-organizing map as initial cluster centers.

**2**: Form k clusters by assigning each node to its closest cluster center.

**3**: Re-compute the cluster centers as the means of all its cluster members.

**4**: Repeat the process from step 2 until the cluster centers no longer change.

K-means was run for different values of *k*, and we chose the optimal partition of the self-organizing map, validated by the Davies-Bouldin index [[1](#_ENREF_1)], so that distances within clusters were minimized and distances between clusters were maximized. The Davies-Bouldin index minimizes the expression:

 $\frac{1}{C}\sum\_{i=1}^{C}max\_{j}\left(\frac{S\_{i}+ S\_{j}}{M\_{ij}}\right)$

where *C* is the number of clusters, *Si* is the dispersion of cluster *i* defined in terms of mean squared distance from the cluster center, and *Mij* is the distance between the centers of cluster i and *j* [[2](#_ENREF_2)]. Thus, the optimal partition implies that, by grouping families based on similarity of family history, a family is then more similar to any family belonging to the same cluster than with any other family in a different cluster.

Finally, *k* cluster-wide familial risk categories were revealed by finding the prototype vectors corresponding to the *k* cluster centers. A cluster-wide familial risk category characterizes each family of a cluster by summarizing the global characteristics of cancer aggregation of all families in that cluster. It is essentially the mean vector of all codebook vectors associated to a cluster.

**Appendix C: Distance measure for similarity of familial aggregation**

Central to every cluster algorithm is a metric for measuring distance (or similarity) between objects. Euclidean distance

 $d\left(x\_{1},x\_{2}\right)=\sqrt{\left‖x\_{1}\right‖^{2}+ \left‖x\_{2}\right‖^{2}-2x\_{1}^{'}x\_{2}}$

is the default distance measure for most clustering algorithm, including the Self-organizing map and k-means. One limitation of the Euclidean distance is that it does not discriminate features which are present in one vector but absent in another vector [[3](#_ENREF_3)], making it incapable of recognizing similarity of familial aggregation in epidemiological sense. For example, we have 3 families (*a*, *b* and *c*) and each family is represented by a 4-dimensional familial aggregation vectors featuring 4 cancers:

*a* = (0, 0.1, 0.1, 0 )

*b* = (0.1, 0, 0, 0.1)

*c* = (0.1, 0.2, 0.2, 0.1)

There is no common aggregating cancer between family *a* and family *b*, but there are two common aggregating cancers between family *a* and family *c*. In terms of familial aggregation, families sharing no common aggregating cancers should not be considered similar. Therefore, family *a* should be more similar to family *c* than family *b*, but Euclidean distance delivers count-intuitive result, *d*(*a*,*b*)=0.2 and *d*(*a*,*c*)=0.2, indicating that family *a* is equally similar to family *b* and family *c*.

To overcome limitations of the Euclidean distance, we adopted the extended Jaccard distance [[3](#_ENREF_3)] as an alternative.

 $d\left(x\_{1},x\_{2}\right)=1- \frac{x\_{1}^{'}x\_{2}}{\left‖x\_{1}\right‖^{2}+ \left‖x\_{2}\right‖^{2}-x\_{1}^{'}x\_{2} }$

It is bounded between 0 and 1 with 0 representing perfect match and 1 representing there is no similarity at all. The extended Jaccard distance overcomes limitation of the Euclidean distance by comparing features shared by both vectors against features present in just either one of the two vectors. As such, it will measure similarity of familial aggregation in a more epidemiological sensible manner, by comparing weights of aggregation cancers shared by two families against weights of cancers aggregating in just either one of two families, indicating that, *d*(*a*,*b*)=1 and *d*(*a*,*c*)=0.5, suggesting that family *a* is more similar to family *c* than family *b.*