Supplementary methods

EACCD

The EACCD [1] is an unsupervised learning algorithm designed to partition patients according to the survival time, censoring status as well as measurements on a sequence of selected categorical variables. Development has targeted its application and improvement [2,3,4,5,6,7]. The algorithm consists of 3 steps: defining initial dissimilarities (in terms of the difference between survival functions) between combinations, computing learned dissimilarities, and performing hierarchical clustering of the combinations. Below is one version of the algorithm that utilizes the two-phase Partitioning Around Medoids algorithm (PAM) [8].

Given a collection of combinations $\{C_1, C_2, \cdots, C_n\}$ and nonnegative weights $w_1, w_2, \ldots, w_n$ with $\sum_{k=1}^{n} w_k = 1$.

1. Define the initial dissimilarity $dis_0(C_i, C_j)$ for any pair C_i and C_j.
2. For each k with $1 \leq k \leq n$, apply the two-phase PAM and the initial dissimilarities in Step 1 to partition combinations into k clusters, and define $\delta_k(i,j) = 1$ if C_i and C_j are not assigned into the same cluster and $\delta_k(i,j) = 0$ otherwise. Compute the learned dissimilarity $dis(C_i, C_j) = \sum_{k=1}^{n} w_k \delta_k(i,j)$.
3. Perform hierarchical clustering to cluster the combinations by using $dis(C_i, C_j)$.

In Step 1, the initial dissimilarity can be defined as the value of a test statistic, such as the log-rank test statistic, Gehan-Wilcoxon test statistic, and Tarone-Ware test statistic. When the sizes of combinations are big, better initial dissimilarities can be defined by effect-size based measures, such as hazard ratios and Mann-Whitney parameters [5,7].

Step 2 utilizes initial dissimilarities in Step 1 and an ensemble process to compute the learned dissimilarities, which are more data driven than the initial dissimilarities. The two-phase PAM is used in the ensemble process to partition combinations. The results from PAM are then combined to generate the learned dissimilarity, which is simply the weighted percentage of the times two combinations are not placed into the same cluster by the PAM algorithm. One simple selections of weights is $w_k = 1/kw$ with $w = 1/1 + 1/2 + \ldots + 1/n$ for $k = 1, 2, \ldots, n$. In early versions of EACCD, learned dissimilarities were obtained by averaging the results from many runs of partition methods, which could take a long time to complete if a huge number of runs were used. In contrast, Step 2 above only requires to run PAM n times, a number equal to the number of combinations.

Step 3 clusters the combinations by the learned dissimilarities from Step 2 and a linkage method. Single linkage, average linkage, complete linkage, minimax linkage [9,10], or other agglomerative hierarchical clustering methods may be used in this step. The primary output is a

dendrogram that provides a graphical summary of patients' survival based on the levels of prognostic factors or variables.

In this paper, the initial dissimilarity in Step 1 is based on the Mann-Whitney parameter described below; the weights in Step 2 are chosen to be $w_1 = \ldots = w_K = 1/n$; and the complete linkage method is used in Step 3.

Mann-Whitney parameter

The Mann-Whitney parameter arises from the widely used Mann-Whitney test [11] that examines whether one of two random variables is stochastically larger than the other. Let T_1 and T_2 denote the variables of survival time for patients from population 1 (with survival function $S_1(t)$) and population 2 (with survival function $S_2(t)$), respectively. The Mann-Whitney parameter is defined as $P(T_1 > T_2)$, the probability that a randomly chosen patient from population 1 has a longer survival time than a random chosen patient from population 2. The difference between the Mann-Whitney parameter and 0.5 suggests a difference between $S_1(t)$ and $S_2(t)$. Efron proposed to use $\widehat{D} = -\int_0^\infty \hat{S}_1(\mathrm{t})\, \hat{S}_2(\mathrm{t})\, dt$ to estimate the Mann-Whitney parameter for censoring data, where $\hat{S}_1(\mathrm{t})$ and $\hat{S}_2(\mathrm{t})$ are Kaplan-Meier estimates of $S_1(t)$ and $S_2(t)$, respectively [12]. However, Efron's estimator requires that both $\hat{S}_1(\mathrm{t})$ and $\hat{S}_2(t)$ drop to 0 at the maximum study time (the longest following-up time) and is rather unstable when censoring occurs due to incomplete follow up [13]. To overcome this problem, Wang et al. proposed to 1) use the conditional probability $P(T_1 > T_2 \,|T_1 \leq \tau \; or \; T_2 \leq \tau)$ instead of Mann-Whitney parameter to study the difference between $S_1(t)$ and $S_2(t)$; and 2) use $\widehat{D}_c = \frac{-\int_0^\tau \hat{S}_1(\mathrm{t})\, \hat{S}_2(\mathrm{t})\, dt}{1-\hat{S}_1(\tau)\, \hat{S}_2(\tau)}$ to estimate $P(T_1 > T_2 \,|T_1 \leq \tau \; or \; T_2 \leq \tau)$ [7]. Note that the estimator $\widehat{D}_c$ only requires the survival information up to a time point τ.

In this paper, $|\widehat{D}_c - 0.5|$ is used to compute the initial dissimilarity in survival between two combinations, with τ set to be the maximum possible time by which the Kaplan-Meier estimates of the survival of all combinations can be calculated.

1. Chen D, Xing K, Henson D et al. Developing prognostic systems of cancer patients by ensemble clustering. Biomed Res Int 2009; 2009.
2. Qi R, Zhou S. A comparative study of algorithms for grouping cancer data. In IAENG International Conference on Data Mining and Applications 2014.
3. Qi R, Wu D, Sheng L et al. On an ensemble algorithm for clustering cancer patient data. BMC Syst Biol 2013; 7(4): S9.
4. Chen D, Hueman MT, Henson DE, Schwartz AM. An algorithm for expanding the TNM staging system. Future Oncol 2016; 12(8): 1015-24.

5. Wang H, Chen D, Hueman MT et al. Clustering big cancer data by effect sizes. In Second IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies 2017 Jul 17 (pp. 58-63). IEEE Press.
6. Hueman MT, Wang H, Yang CQ et al. Creating prognostic systems for cancer patients: A demonstration using breast cancer. Cancer Med 2018; 7(8): 3611-21.
7. Wang H, Hueman M, Pan Q et al. Creating Prognostic Systems by the Mann-Whitney Parameter. In 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies 2018 Sep 26 (pp. 33-39). IEEE Press.
8. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis: John Wiley & Sons, Hoboken, 2009.
9. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: prediction, inference and data mining: Springer-Verlag, New York, 2009.
10. Bien J, Tibshirani R. Hierarchical clustering with prototypes via minimax linkage. J AM STAT ASSOC 2011; 106(495): 1075-84.
11. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947; 18: 50-60.
12. Efron B. The two sample problem with censored data. In M. Lucien, Le Cam and Jerzy N (eds) Fifth Berkeley Symposium on Mathematical Statistics and Probability, Statistical Laboratory of the University of California, Berkeley, June 21–18 July 1965 and 27 December 1965–7 January 1966, p.666. Berkeley, Calif: University of California Press 1967.
13. Brentnall AR, Cuzick J. Use of the concordance index for predictors of censored survival data. Stat Methods Med Res 2018; 27(8): 2359-73.