

Claims

1. A computerized method for identifying one or more genes, or pairs of genes, selected from a set of N genes, which are statistically associated with prognosis of a potentially fatal medical condition, the method employing test data which, for each subject k of a set of K subjects suffering from the medical condition, indicates (i) a survival time of subject k , and (ii) for each gene i , a corresponding gene expression value $y_{i,k}$ of subject k ;

the method comprising:

(A) for each of a plurality of genes or pairs of genes, performing the sub-steps of:

- (a) partitioning the K subjects into two subsets using cut-off values, (b) computationally fitting a Cox proportional hazard regression model to the corresponding test data of at least one of the subsets of subjects;
- (c) determining whether a proportionality assumption of the Cox proportional hazard regression model is satisfied using the Bayesian Information Criterion;
- (d) if the proportionality assumption is satisfied, fitting one or more additional proportional hazard regression models to the data-set;
- (e) if the proportionality assumption is not satisfied, fitting at least one hazard regression model which does not satisfy a proportionality assumption to the data-set;
- (f) selecting, from the fitted hazard regression models, a hazard regression model having best quality of fit based on the Bayesian Information Criterion; and
- (g) obtaining a significance value indicative of prognostic significance of the gene or gene pair using the selected hazard regression model;

wherein when operation (A) is performed for each of the plurality of pairs of genes (i, j with $i \neq j$) of the set of N genes, operation (A) comprises: generating a respective plurality of trial cut-off values of $c_{i,j}$, and for each of the trial cut-off values:

partitioning the K subjects into two subsets according to whether $\log(y_{i,k}) - \log(y_{j,k})$ is respectively above or below the trial cut-off value $c_{i,j}$;

performing sub-operations (b)-(f); and

obtaining a significance value indicative of prognostic significance of the gene pair i,j using the selected hazard regression model; and

for each of the pairs of genes, identifying the trial cut-off value $c_{i,j}$ for which the corresponding significance value indicates the highest prognostic significance for the gene pair i,j ; and

(B) identifying one or more of the genes or pairs of genes for which the corresponding significance values have the highest prognostic significance.

2. A method according to claim 1 in which, in sub-step (e) the method includes fitting a stratified Cox model to the data-set.

3. A computerized method according to any preceding claim wherein the medical condition is cancer, and the expression levels are from samples of respective tumours in the K subjects.

4. A method according to claim 3 wherein the medical condition is breast cancer.

5. A method according to any preceding claim in which the survival time is survival time until mortality.

6. A method according to any preceding claim in which the survival time is a survival time without metastasis of a cancer.

7. A method according to any preceding claim further comprising:

(i) for each of the one or more identified genes, obtaining corresponding gene expression values of a first subject; and

- (ii) obtaining information about the first subject in relation to the potentially-fatal medical condition using the obtained gene expression values, the selected hazard regression model and the cut-off values.
- 8. A method according to any preceding claim further comprising:
 - (i) for each of the one or more identified pairs genes, obtaining corresponding gene expression values of a first subject; and
 - (ii) obtaining information about the first subject in relation to the potentially-fatal medical condition using the obtained gene expression values, the selected hazard regression model and the identified trial cut-off value.
- 9. A method according to claim 7 or claim 8 in which said information is a prognosis for the first subject who is suffering from the medical condition, a susceptibility of the first subject to the medical condition, a prediction of the recurrence of the medical condition, or a recommended treatment for the medical condition.
- 10. A method according to any preceding claim, wherein sub-step (e) further includes fitting an additional non-proportional hazard regression model to the data-set.
- 11. A method according to any preceding claim further including performing step (A) for each of a plurality of genes and pairs of genes of the set of N genes.
- 12. A computer system arranged to perform a method according to any preceding claim.

13. A data carrier, such as a tangible data carrier, carrying software readable by a computer to cause the computer to perform a method according to any of claims 1 to 11.