Statistical tests for heterogeneity of clusters and composite endpoints

Anthony J. Webster

Nuffield Department of Population Health, Big Data Institute, Old Road Campus, University of Oxford, Oxford, OX3 7LF, UK.

Abstract

Clinical trials and epidemiological cohort studies often group similar diseases together into a composite endpoint, to increase statistical power. A common example is to use a 3-digit code from the International Classification of Diseases (ICD), to represent a collection of several 4-digit coded diseases. More recently, data-driven studies are using associations with risk factors to cluster diseases, leading this article to reconsider the assumptions needed to study a composite endpoint of several potentially distinct diseases. An important assumption is that the (possibly multivariate) associations are the same for all diseases in a composite endpoint (not heterogeneous). Therefore, multivariate measures of heterogeneity from meta-analysis are considered, including multi-variate versions of the I^2 and Q statistics. Whereas meta-analysis offers tools to test heterogeneity of clustering studies, clustering models suggest an alternative heterogeneity test, of whether the data are better described by one, or more, clusters of elements with the same mean. The assumptions needed to model composite endpoints with a proportional hazards model are also considered. It is found that the model can fail if one or more diseases in the composite endpoint have different associations. Tests of the proportional hazards assumption can help identify when this occurs. It is emphasised that in multi-stage diseases such as cancer, some germline genetic variants can strongly modify the baseline hazard function and cannot be adjusted for, but must instead be used to stratify the data.

Introduction

It is common for epidemiological studies [1] and clinical trials [2–4] to group similar diseases together into a cluster of diseases, providing more total cases, and more statistical power to detect associations. This procedure has been essential since the pioneering epidemiological studies of John Graunt in the 1600s [5], providing sufficient cases to allow meaningful statistical study, while attempting to ensure that clustered diseases have a similar etiology. Today an increasingly detailed biological understanding informs the clustering of diseases. One common approach is to use the heirarchical classification of the International Classification of Diseases (ICD) as a guide for which diseases to group [6]. The ICD system clusters increasingly detailed disease descriptions into larger clusters with similar disease etiology, often allowing larger clusters to be used in epidemiological studies. In practice, diseases are selected by a clinician to help ensure that only diseases with a clearly defined and common etiology are clustered together [1]. With data-driven clustering studies increasingly being used to identify potential composite endpoints [1, 7–11], it seems appropriate to review the assumptions needed to study a cluster of diseases, and the existing arguments for and against doing so in clinical trials.

In composite endpoints, or diseases clustered using shared risk factors [1], a necessary requirement is that risk-factor associations are the same (homogeneous). A similar requirement arose in meta-analysis, and led to the development of heterogeneity tests [12]. The multivariate versions of these tests, their origin, and their application to assess the heterogeneity of risk-factor associations within a composite endpoint or a clustering study is described. In the process, an alternative clustering-based heterogeneity test is suggested that offers a different perspective to the conventional Q and I^2 statistics [13, 14], that are commonly used in meta-analyses [12].

The proportional hazards modelling is ubiquitous in medical research and statistical epidemiology, so the implicit assumptions needed when using it to model a composite endpoint are explored. Heterogeneity of disease associations with risk-factors in a composite endpoint, are sufficient to cause the proportional hazards assumptions to fail. Therefore it is necessary for tests of the proportional hazards assumption to be satisfied, and these can be used when tests of heterogeneity are inappropriate due to insufficient data for example. From a statistical modelling perspective, the key requirement is that risk-factor associations are the same for diseases within a composite endpoint, baseline incidence rates can be arbitrarily different. (Although as will be discussed in the context of clinical trials, insufficient cases of a disease that prevent tests of heterogeneity, should only be included if there are strong prior reasons to expect the same risk-factor associations as for the other diseases in the endpoint.) The conceptual multi-stage model of diseases [15] such as cancer [16–18] and motor neuron disease [19–21], is used to highlight some additional implicit assumptions that are needed when the proportional hazards methodology is used.

Composite endpoints in clinical trials

It is worth briefly recapping the benefits and concerns about using composite endpoints in clinical trials.

Reasons to consider composite endpoints

There are several good reasons to use composite endpoints:

i) Diseases are implicitly defined as a composite endpoint of symptoms and biological measurements. (More precise subgroups sometimes have different behaviours [9, 10, 22, 23].)

ii) Statistical power - If a drug targets a pathway that modifies the risk of several otherwise distinct diseases, or if a disease is a symptom of onset of more serious disease, then it makes sense to study these together to increase statistical power.

The importance of both i) and ii) are apparent in the tables of disease studied by John Graunt in the 1600s [5], where it is difficult to

compare diseases when cases are rare or erratic. Weaker arguments to use composite endpoints include:

iii) Interest is in avoiding any negative disease outcomes, not just particularly severe ones.

iv) Interest is in testing the potential influence of a potential drug or risk factor on a wide range of diseases.

These arguments iii) and iv) are more easily criticised as "fishing", casting a wider net to try and catch more diseases that may be influenced by a new drug without strong prior reason to do so, and increasing the risk of false positive associations.

Criticisms of composite endpoints

There are several serious criticisms and limitations of using composite endpoints, some of which are easily corrected, but others that appear unavoidable:

a) Early criticisms of composite endpoints were, to quote Cordoba et al. [24], that "Components are often unreasonably combined, inconsistently defined, and inadequately reported.".

b) Endpoints can be of different importance to patients. This has led several authors to argue for weighting of endpoints based on importance to patients [2, 25, 26].

c) Infrequent disease - Insufficient cases of a disease in a composite endpoint can make it impossible to test if effect sizes are comparable, e.g. due to unreliable confidence intervals.

d) Similar effect sizes - If effect sizes are dissimilar, then it is impractical to meaningfully interpret results. If disease risk is being modified through a shared disease pathway, then we might expect a similar proportional reduction in disease risk.

Criticisms in a) can be avoided by careful study design, specification, and reporting. The concerns in c) limit studies to diseases with sufficient cases of each distinct disease to allow meaningful tests of similar effect sizes. A more nuanced concern is b), the argument that endpoints should be of equal importance to different patients. To develop understanding and treatments, it is most important to identify endpoints with a shared disease pathway. Whereas diseases sharing this pathway could have very different relevance to patients, diseases should only be included if they have the same underlying cause.

Despite the limitations, points (i) and (ii) emphasise that composite endpoints are often reasonable, or necessary, and in these cases it is important to understand the assumptions being made when we study them. Research that is intended to cluster diseases by common disease pathways will also lead to newly hypothesised composite endpoints, and it is helpful to establish tests and understand the assumptions that disease clusters are consistent with.

Heterogeneity of associations

Heterogeneity tests are widely used in meta-analyses, and are intended to assess whether the reported associations are the same in several different studies [12]. For composite endpoints we wish to assess whether one or more associations are the same for all diseases in a composite endpoint. The method's multivariate generalisation that is needed for clustering studies such as Ref. [1], is described below. Multivariate heterogeneity tests were originally developed for meta-analyses [27–29], for which case a random effects model is often more suitable. A discussion of the multivariate test for meta-analyses in a random effects model, can be found in [30]. Here we explain the basis for the (fixed effects) multivariate heterogeneity test, that is most relevant to composite endpoints. Later a test that originates from clustering studies is suggested, that offers an alternative approach with some advantages over conventional Q and I^2 statistics [12].

The null hypothesis (in a fixed effects model), is that all diseases in a composite endpoint have the same associations with one or more parameters, such as a drug, or a collection of potential risk factors. These might be a subset of associations, with potential confounders adjusted for, and subsequently removed by marginalisation [1]. Consider m composite endpoints (or clusters of diseases), labelled by g. Under the null hypothesis of the same associations for diseases in a composite endpoint, labelled $i = 1$ to $i = n_g$,

$$X_i \sim N\left(\mu_g, \Gamma_i^{-1}\right) \tag{1}$$

where $\Gamma_i^{-1} = \Sigma_i$ is the covariance (Γ_i is the precision matrix), and μ_g are the (unknown) associations that we are estimating, that are assumed to be the same for all diseases in the composite endpoint. Eq. 1 requires [31],

$$\left(X_i - \mu_g\right)^T \Gamma_i \left(X_i - \mu_g\right) \sim \chi_p^2 \tag{2}$$

where p is the dimension. Therefore because the sum of n_g random variables that are individually χ_p^2 distributed is $\chi_{pn_g}^2$,

$$\sum_{i=1}^{n_g} \left(X_i - \mu_g\right)^T \Gamma_i \left(X_i - \mu_g\right) \sim \chi_{pn_g}^2 \tag{3}$$

For $p = 1$ this has,

$$\sum_{i=1}^{n_g} \left(\frac{X_i - \mu_g}{\sigma_i}\right)^2 \sim \chi_{n_g}^2 \tag{4}$$

for standard deviations σ_i. Because μ_g is unknown and must be estimated, the test statistic is modified, as explained next.

Using Bayes theorem with either a flat or normal prior for the mean μ_g, a cluster of diseases with covariances $\{\Gamma_i\}$ and the same mean μ_g, have [32],

$$\mu_g \sim N\left(\tilde{\mu}_g, \tilde{\Lambda}_g^{-1}\right) \tag{5}$$

where,

$$\tilde{\mu}_g = \left(\sum_i \Gamma_i\right)^{-1} \left(\sum_i \Gamma_i X_i\right) \tag{6}$$

and,

$$\tilde{\Lambda}_g = \left(\sum_i \Gamma_i\right) \tag{7}$$

where if a normal prior is used then the sum over i includes the prior's mean μ_0 and covariance Λ_0, and the sum is from $i = 0$ to $i = n_g$. For a flat prior, the sum is from $i = 1$ to $i = n_g$. The

subscripts g allow the discussion to include more than one cluster of diseases, as was considered in Webster et al. [1]. Here unless stated otherwise, we consider a single composite endpoint, and the subscript g could be omitted.

As a result of Eq. 5, we have,

$$\left(\mu_g - \tilde{\mu}_g\right)^T \tilde{\Lambda}_g \left(\mu_g - \tilde{\mu}_g\right) \sim \chi^2_p \tag{8}$$

These observations, together with Eq. 1, can be used to derive a multivariate test for the assumption that the normal distributions have the same mean. Appendix A shows that,

$$\begin{aligned} &\textstyle\sum_i \left(X_i - \tilde{\mu}_g\right)^T \Gamma_i \left(X_i - \tilde{\mu}_g\right) \\ &= \textstyle\sum_i \left(X_i - \mu_g\right)^T \Gamma_i \left(X_i - \mu_g\right) - \left(\mu_g - \tilde{\mu}_g\right)^T \tilde{\Lambda}_g \left(\mu_g - \tilde{\mu}_g\right) \end{aligned} \tag{9}$$

Using this with Eqs. 3 and 8,

$$\sum_{i=1}^{n_g} \left(X_i - \tilde{\mu}_g\right)^T \Gamma_i \left(X_i - \tilde{\mu}_g\right) \sim \chi^2_{(n_g-1)p} \tag{10}$$

The left side of Eq. 10 is the Q statistic. It provides a test for the assumption that the (approximately) normally distributed estimates $\{X_i\}$ have the same mean, as is assumed for a composite endpoint or cluster of diseases. For $p = 1$, these expressions give the well known inverse variance weighted heterogeneity test, that are regularly used in meta analyses and 2-sample Mendelian randomisation studies [12, 33].

For the situation described in Webster et al. [1], the aim is to assess the goodness of fit for a clustering of diseases. In effect, there is a set of composite endpoints being considered, and the task is to determine the optimum split of diseases into composite endpoints with similar risk-factor associations. For this situation, Eq. 10 is modified to sum over all m clusters, and the Q statistic becomes,

$$\sum_{g=1}^{m} \sum_{i \in C_g} \left(X_i - \tilde{\mu}_g\right)^T \Gamma_i \left(X_i - \tilde{\mu}_g\right) \sim \chi^2_{(n-m)p} \tag{11}$$

where we used $\sum_{g=1}^{m}(n_g - 1)p = (n - m)p$, and C_g is the set of diseases in cluster g (composite endpoint g).

The results above correspond to a fixed effects model where the data are assumed to have the same mean, as opposed to the means being sampled from an underlying distribution (a random effects model). A random effects model samples each study's mean μ_i, with $\mu_i \sim N(\mu_0, \Gamma_0^{-1})$, and assumes measured estimates X_i have $X_i \sim N(\mu_i, \Gamma_i^{-1})$. Marginalising out μ_i gives $X_i \sim N(\mu_0, \Gamma_0^{-1} + \Gamma_i^{-1})$, that replaces Eq. 1. Assuming that the data belong to a single cluster, then Eq. 5 becomes $\mu_0 \sim N(\tilde{\mu}_0, \tilde{\Lambda}_0^{-1})$, and Eqs. 6 and 7 become,

$$\tilde{\mu}_0 = \left(\sum_i \left(\Gamma_0^{-1} + \Gamma_i^{-1}\right)^{-1}\right)^{-1} \left(\sum_i \left(\Gamma_0^{-1} + \Gamma_i^{-1}\right)^{-1} X_i\right) \tag{12}$$

and

$$\tilde{\Lambda}_0 = \sum_i \left(\Gamma_0^{-1} + \Gamma_i^{-1}\right)^{-1} \tag{13}$$

as in Jackson et al. [30] (where $\tilde{\Lambda}_0$ is the inverse of the covariance). To calculate the Q and I^2 statistics using Eq. 10, Γ_i is replaced by $(\Gamma_0^{-1} + \Gamma_i^{-1})^{-1}$, $\tilde{\mu}_g$ with $\tilde{\mu}_0$, and n_g is the total number of studies. The above arguments and results could be modified to consider a random effects model with different priors for each cluster.

Assessing heterogeneity within meta-analyses

The classical measure of heterogeneity uses the Q statistic, that was derived above in a multi-variate context. The I-square statistic [13, 14] is closely related to the Q statistic [12], and in the notation above, is for Eq. 11,

$$I^2 = \frac{Q - (n - m)p}{Q} \times 100 \tag{14}$$

where Q is the left side of Eq. 11, and the factor of 100 is conventionally used to express I^2 as a percentage. Unlike in figures 1 and 2, I^2 is usually set to zero if its evaluation is negative. The equivalent expression for a single cluster that uses the left side of Eq. 10 for Q, would replace the number of degrees of freedom $(n - m)p$, with $(n_g - 1)p$ in Eq. 14. The I^2 statistic replaces a test with a more nuanced measure of heterogeneity that is particularly useful when some heterogeneity is expected, but it no longer provides an objective statistical test.

This article was originally motivated by the need to better understand and characterise the results from clustering studies, such as assessing the heterogeneity of the underlying clusters. However, clustering studies are becoming increasingly rigorous (Figures 1 and 2), and can offer an alternative approach to test the heterogeneity of composite endpoints. Webster [32] provides a log-likelihood for clustering normally distributed data such as that from maximum likelihood estimates, that assumes equal means in each cluster. The log-likelihood is maximised to determine whether the data fit more naturally in one or more clusters. This provides an intuitive and objective assessment of heterogeneity, that in the examples shown in figures 1 and 2, was more lenient than a statisical test using Q. Conventional heterogeneity tests estimate the mean effect size across several groups, and then assess or test, whether the mean effect size is consistent with all the studies. In contrast, when assessing clusterings of data with the model in Webster [32], equal means are assumed in all clusters, and minimising the log-likelihood determines whether the study data are best explained by one, or more clusters. This objective test is different to asking whether there are statistical differences between data, and is more tolerant of heterogeneity. It is easy to incorporate prior information into the approach, such as a weak normal prior for zero effect size. The selection of priors in a (1-dimensional) Bayesian meta-analysis is discussed in Ott et al. [34]. In principle, clustering methods could allow the identification of statistically distinct subgroups or outliers. This and the relative merits of the approach need considering in greater detail elsewhere, but it appears to offer a potentially useful alternative to conventional measures of heterogeneity.

Composite endpoints in proportional hazards studies

Composite endpoints are frequently studied with proportional hazards models. This section explores the assumptions that are needed for the model to be correct, and for its estimates to be meaningful.

Firstly consider m diseases in a composite endpoint that may be influenced by the same factors or processes, but are otherwise independent of each other. The survival probability is determined from the probability of an individual surviving all m diseases until time t. If $S_j(t)$ is the probability of surviving disease j to time t, and $S(t)$ is the probability of surviving all m independent diseases,

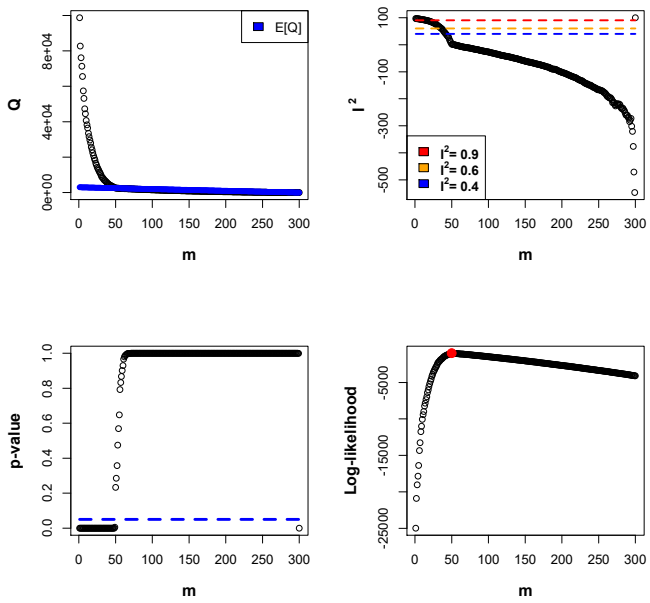

Figure 1: For a flat prior and normally distributed test data [32], the Q-statistic and log-likelihood correctly identify the 50 clusters of test data with 2 to 14 members each [32]. Interestingly, the rate of reduction of I^2 slowed after the 50 clusters were identified.

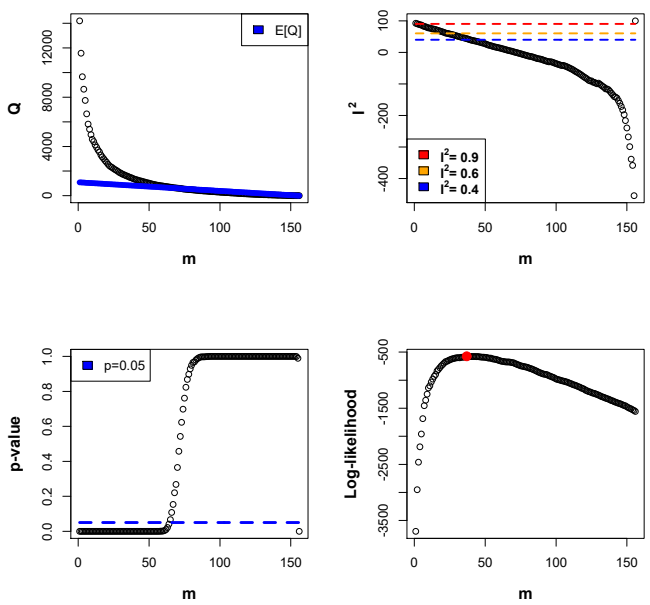

Figure 2: For a flat prior and UK Biobank disease data [1], the clustering log-likelihood of Webster [32] is minimised at $m = 37$, where $I^2 \simeq 50\%$. A statistical heterogeneity test using Q fails for less than 63 clusters. With a normal prior centred on zero, I^2 reduces and the maximum in log-likelihood moves to $m \sim 25$, and was fairly insensitive to the prior's covariance [32].

then [15],

$$S(t) = \Pi_{j=1}^{m} S_j(t) \tag{15}$$

Writing $S_j(t)$ in terms of its hazard function $h_j(t)$, with $S_j(t) = \exp\{-\int_0^t h_j(s)ds\}$, then Eq. 15 requires,

$$S(t) = \exp\left\{-\int_0^t \sum_{j=1}^{m} h_j(s)ds\right\} \tag{16}$$

giving the hazard function $h(t)$ for a group of diseases as [15],

$$h(t) = \sum_{j=1}^{m} h_j(t) \tag{17}$$

The proportional hazards model takes $h(t) = e^{\beta X_i} h_0(t)$ for an individual i with covariates X_i, and assumes that $h_0(t)$ is the same for all individuals in a population. If the assumption holds for each disease in a cluster or composite endpoint, then we can write,

$$h(t) = \sum_{j=1}^{m} e^{\beta_j X_i} h_{0j}(t) \tag{18}$$

If all the diseases have the same risk-factor associations β, then,

$$h(t) = e^{\beta X_i} \sum_{j=1}^{m} h_{0j}(t) \tag{19}$$

and the proportional hazards assumption will remain true for the cluster of diseases, with all the time-dependence in a single factor that will be the same for all individuals. However if any disease k has $\beta_k \neq \beta$, then,

$$h(t) = e^{\beta_k X_i} h_{ok}(t) + e^{\beta X_i} \sum_{j \neq k} h_{0j}(t) \tag{20}$$

and the proportional hazards assumption fails. Note that this is the case if the assumption fails for any single covariate. The observations described by Eqs. 15-20 have the following implications:

1. For the proportional hazards assumption to hold, it is necessary that all clustered diseases have the same associations.
2. Counter-intuitively, the baseline hazard for each individual disease could be very different with diseases prevalent at very different ages, but the cluster of diseases can still satisfy the proportional hazards assumption.
3. If the proportional hazards model fails, it could be due to one or more of the clustered diseases having different risk associations - it need not necessarily be caused by different baseline hazards $h_{0j}(t)$ in the underlying population.

Points 1 and 2 emphasise that diseases must be clustered on the basis of similar underlying etiology, and that incidence rates are irrelevant and potentially misleading for determining whether diseases can be considered as a single cluster and modelled with a proportional hazards model. Point 3 indicates that failure of the proportional hazards model, as identified by a statistical test using the Schoenfeld residuals for example [35], could be due to one or more diseases having different risk factors. This provides a consistency test for whether a cluster of diseases have the same risk factors, without fitting and comparing each disease individually, which may not always be possible.

General remarks

If statistical tests are consistent with the proportional hazards assumption, then it can be reported that the diseases in the composite endpoint are consistent with having the same associations (within

a proportional hazards model). This does not ensure that diseases have the same associations, only that there is no strong statistical evidence for them being different. Importantly, as was noted in the context of clinical trials, if the number of cases of a disease are few then they are unlikely to have much influence on a test result. In those cases the disease should only be included if there are strong prior (usually biological), reasons to include them.

The discussion above has considered the commonly used proportional hazard model. It has not considered the influence of model misspecification or the censoring distribution [36], or any newer methods for modelling composite endpoints that are being developed [37, 38].

Multi-stage disease processes

Although Eq. 19 allows different baseline hazards for the diseases in a composite endpoint, the proportional hazards methodology requires that the resulting baseline hazard $\sum_j h_{0j}(t)$, is the same for all individuals (or individuals in a strata of a stratified analysis). Some rates of disease incidence, such as Cancers [16–18] and Amyotrophic Lateral Sclerosis [19–21], can be described by multi-stage disease processes [15], where one or more rate-limiting steps to disease may be skipped by a germline genetic variation. This will produce different baseline hazards for individuals with the genetic change. Importantly however, such changes will qualitatively modify the incident rate to a different power of time [15–21], and cannot be adjusted for in a proportional hazards model. Instead, the variant should be used to stratify the data, so that different strata have differing baseline hazards.

Conclusions

Composite endpoints are intrinsic to how we define and study disease. Since John Graunt's studies [5], there has always been a trade-off between disease definitions that are sufficiently specific to distinguish different underlying disease processes, but also sufficiently broad to allow a meaningful statistical study. This is particularly apparent in clinical trials and epidemiological studies where data are costly or unavailable. Large population datasets with detailed genetic and biological information are providing a new data-driven source of composite endpoints, by identifying composite endpoints with potential shared underlying causes.

Statistical methods can assess whether a composite endpoint is consistent with its assumed properties, such as testing its constituent diseases for heterogeneity among their disease-risk associations. Heterogeneity is conventionally tested with a Q or I^2 statistic [12]. An alternative clustering-based approach, is to assume that diseases are in one or more clusters with equal associations, and test if the log-likelihood for the model [32] is minimised by one, or more clusters. This objective test can incorporate a prior, and examples suggest it is more lenient than a Q test, with the disease clustering data of Webster et al. [1] minimising the log-likelihood when $I^2 \simeq 50\%$. The merits of this approach for applications such as meta-analyses, will need exploring in greater detail elsewhere.

When proportional hazard models are used, then heterogeneity of associations in a composite endpoint will cause the proportional hazards assumption to fail. However, the statistical model does allow arbitrarily different baseline incidence rates for the diseases in the composite endpoint (although each baseline hazard h_{0j} must be the same for all individuals in the studied population). A test for the assumption of proportional hazards, provides a necessary consistency test when heterogeneity tests are impractical or impossible; although the proportional hazards assumption could fail for other reasons.

The multi-stage model of disease emphasises that germline genetic variants that modify the number of rate-limiting steps prior to disease, cannot be adjusted for within a proportional hazards model but must be used to stratify the data. This remark holds more generally, regardless of whether a disease is part of a composite endpoint or studied alone.

Appendix: Multivariate inverse variance weighted sum of squares

Note that the covariances and their inverses are symmetric, and expand,

$$\begin{aligned}
&\textstyle\sum_i (X_i - \mu_g)^T \Gamma_i (X_i - \mu_g) \\
&= \textstyle\sum_i ((X_i - \tilde{\mu}_g) - (\mu_g - \tilde{\mu}_g))^T \Gamma_i ((X_i - \tilde{\mu}_g) - (\mu_g - \tilde{\mu}_g)) \\
&= \textstyle\sum_i (X_i - \tilde{\mu}_g)^T \Gamma_i (X_i - \tilde{\mu}_g) + \sum_i (\mu_g - \tilde{\mu}_g)^T \Gamma_i (\mu_g - \tilde{\mu}_g) \\
&-2 \textstyle\sum_i (\mu_g - \tilde{\mu}_g)^T \Gamma_i (X_i - \tilde{\mu}_g) \\
&= \textstyle\sum_i (X_i - \tilde{\mu}_g)^T \Gamma_i (X_i - \tilde{\mu}_g) + (\mu_g - \tilde{\mu}_g)^T \left(\sum_i \Gamma_i\right) (\mu_g - \tilde{\mu}_g) \\
&-2 (\mu_g - \tilde{\mu}_g)^T \left(\left(\textstyle\sum_i \Gamma_i X_i\right) - \left(\sum_i \Gamma_i\right) \tilde{\mu}_g\right)
\end{aligned} \tag{21}$$

If $\tilde{\mu}$ takes the specific form given by Eq. 6, then the terms $\left(\sum_i \Gamma_i X_i\right)$ and $\left(\sum_i \Gamma_i\right) \tilde{\mu}_g$ in the last term of the final line cancel, and the resulting equation can be rearranged to give,

$$\begin{aligned}
&\textstyle\sum_i (X_i - \tilde{\mu}_g)^T \Gamma_i (X_i - \tilde{\mu}_g) \\
&= \textstyle\sum_i (X_i - \mu_g)^T \Gamma_i (X_i - \mu_g) - (\mu_g - \tilde{\mu})^T \tilde{\Lambda}_g (\mu_g - \tilde{\mu})
\end{aligned} \tag{22}$$

which is Eq. 9.

Acknowledgements

This research has been conducted using data from UK Biobank, a major biomedical database, under application number 42583. Anthony Webster is supported by a fellowship from the Nuffield Department of Population Health (NDPH).

References

[1] Webster, A., Gaitskell, K., Turnbull, I., Cairns, B. & Clarke, R. Characterisation, identification, clustering, and classification of disease. *Scientific Reports* **11**, 5405 (2021).

[2] Sankoh, A. J., Li, H. & D'Agostino, R. B., Sr. Use of composite endpoints in clinical trials. *Statistics in Medicine* **33**, 4709–4714 (2014).

[3] Mccoy, C. E. Understanding the Use of Composite Endpoints in Clinical Trials. *Western Journal of Emergency Medicine* **19**, 631–634 (2018).

[4] Palileo-Villanueva, L. M. & Dans, A. L. Composite endpoints. *Journal of Clinical Epidemiology* **128**, 157–158 (2020).

[5] Graunt, C. J. *Natural and Political OBSERVATIONS Mentioned in a following INDEX, and made upon the Bills of Mortality* (Printed by John Martyn, Printer to the Royal Society, at the Sign of the Bell in St. Paul's Church-yard. MDCLXXVI., 1665). Appendix - The table of casualties - Table of Casualties in Economic Writings (vol. 2) by William Petty (1899), between p. 406 and 407.

[6] Organization, W. H. International statistical classification of diseases and related health problems 10th revision (2016). URL `https://icd.who.int/browse10/2016/en`.

[7] Alhasoun, F. *et al.* Age density patterns in patients medical conditions: A clustering approach. *PLOS Computational Biology* **14** (2018).

[8] Zhou, X. *et al.* A Systems Approach to Refine Disease Taxonomy by Integrating Phenotypic and Molecular Networks. *EBioMedicine* **31**, 79–91 (2018).

[9] Guillamet, R. V., Ursu, O., Iwamoto, G., Moseley, P. L. & Oprea, T. Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records. *Health Informatics Journal* **24**, 394–409 (2018).

[10] Seymour, C. W. *et al.* Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* **321**, 2003–2017 (2019).

[11] Kuan, V. *et al.* Data-driven identification of ageing-related diseases from electronic health records. *Scientific Reports* **11** (2021).

[12] Borenstein, M., Hedges, M., J.P.T., H. & Rothstein, H. *Introduction to Meta-Analysis* (Wiley, 2009).

[13] Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539–1558 (2002).

[14] Higgins, J., Thompson, S., Deeks, J. & Altman, D. Measuring inconsistency in meta-analyses. *British Medical Journal* **327**, 557–560 (2003).

[15] Webster, A. Multi-stage models for the failure of complex systems, cascading disasters, and the onset of disease. *PLOS One* **14**, e0216422. (2019).

[16] Nordling, C. A new theory on the cancer-inducing mechanism. *British Journal of Cancer* **7**, 68–72 (1953).

[17] P., A. & R., D. The age distribution of cancer and a multistage theory of carcinogenesis. *British Journal of Cancer* **8**, 1–12 (1954).

[18] Moolgavkar, S. H. Commentary: Multistage carcinogenesis and epidemiological studies of cancer. *International Journal of Epidemiology* **45**, 645–649 (2015).

[19] Ai-Chalabi, A. *et al.* Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. *The Lancet Neurology* **13**, 1108 – 1113 (2014).

[20] Chiò, A. *et al.* The multistep hypothesis of als revisited. *Neurology* **91**, e635–e642 (2018).

[21] Corcia, P. *et al.* In als, a mutation could be worth two steps. *Revue Neurologique* **174**, 669–670 (2018).

[22] Macedo Hair, F. N. F. . B. P., G. Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach. *BMC Infectious Diseases* **19**, 649 (2019).

[23] Kueffner, Z. N. B. M. e. a., R. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Scientific Reports* **9**, 690 (2019).

[24] Cordoba, G., Schwartz, L., Woloshin, S., Bae, H. & Gøtzsche, P. C. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ* **341** (2010).

[25] Duc, A. N. & Wolbers, M. Weighted analysis of composite endpoints with simultaneous inference for flexible weight constraints. *Statistics in Medicine* **36**, 442–454 (2017).

[26] Armstrong, P. W. & Westerhout, C. M. Composite End Points in Clinical Research A Time for Reappraisal. *Circulation* **135**, 2299 (2017).

[27] Kalaian, H. & Raudenbush, S. A multivariate mixed linear model for meta-analysis. *Psychological Methods* **1**, 227–235 (1996).

[28] van Houwelingen, H., Arends, L. & Stijnen, T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* **21**, 589–624 (2002).

[29] Nam, I., Mengersen, K. & Garthwaite, P. Multivariate meta-analysis. *Statistics in Medicine* **22**, 2309–2333 (2003).

[30] Jackson, D., White, I. R. & Riley, R. D. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine* **31**, 3805–3820 (2012).

[31] Hardle, W. & Simar, L. Applied Multivariate Statistical Analysis, 4th Edition. In *Applied Multivariate Statistical Analysis, 4th edition*, 1–580 (2015).

[32] Webster, A. Clustering parametric models and normally distributed data. *arXiv:* 2008.03974v3 (2020).

[33] Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* **28**, 30–42 (2017).

[34] Ott, M., Plummer, M. & Roos, M. How vague is vague? how informative is informative? reference analysis for bayesian meta-analysis. *Statistics in Medicine* .

[35] Klein, J. & Moeschberger, M. Survival analysis. In *Survival Analysis, Techniques for Censored and Truncated Data, Second Edition*, 1–531 (2003).

[36] Wu, L. & Cook, R. J. Misspecification of Cox regression models with composite endpoints. *Statistics in Medicine* **31**, 3545–3562 (2012).

[37] Hara, H. *et al.* Statistical methods for composite endpoints. *EuroIntervention* **16**, E1484+ (2021).

[38] Rauch, G., Kieser, M., Binder, H., Bayes-Genis, A. & Jahn-Eimermacher, A. Time-to-first-event versus recurrent-event analysis: points to consider for selecting a meaningful analysis strategy in clinical trials with composite endpoints. *Clinical Research in Cardiology* **107**, 437–443 (2018).