Comparing t-year absolute risk prediction strategies: the multi-split testing approach

Paul Blanche[1], Mark van de Wiel[2], Jonas B. Nielsen[3] & Thomas A. Gerds[1]

[1] *Dep. of Biostatistics, Univ. of Copenhagen, Denmark.*
E-mail: pabl@sund.ku.dk.
[2] *Dep. of Epidemiology & Biostatistics, VU University, Amsterdam, The Netherlands.*
[3] *Rigshospitalet, Copenhagen University Hospital, Denmark.*

Résumé. L'intérêt croissant pour la médecine personnalisée crée une demande importante de modèle prédictifs. De nombreux modèles statistiques et stratégies ont déjà été discutés pour construire des outils pronostiques. Simultanément, les capacités pronostiques de nombreux facteurs de risques et nouveaux biomarqueurs sont aujourd'hui évalués. En pratique, ceci complique fortement le choix d'une stratégie, parmi les nombreuses possibles, pour construire un modèle prédictif. Leur comparaison objective est une tâche délicate.

Pour comparer deux stratégies de prédiction, une technique couramment utilisée consiste à diviser les données en deux : un "échantillon d'apprentissage", utilisé pour développer les deux outils de prédiction, et un "échantillon test", utilisé pour les comparer. Malheureusement, les conclusions dépendent souvent de la façon dont les données ont été divisées. Van de Wiel et al. (2009) ont récemment proposé une approche par test basée sur de multiples scissions des données. Les avantages de l'approche incluent son implémentation aisée et son universalité, qui permettent de comparer des stratégies de prédiction très diverses. Elle est également générale en ce qui concerne le critère utilisé pour évaluer les capacités pronostiques.

Des extensions aux situations incluant la présence de données censurées et de risques concurrents sont présentées. Nous discutons aussi de nouveaux résultats concernant les hypothèses de la méthode et le contrôle de son erreur de type-I. Une application à la prédiction d'événements cardiovasculaires est présentée. L'objectif est de comparer des stratégies de prédiction basées sur des électrocardiogrammes. Les données d'une cohorte Danoise de grande taille sont analysées ($n = 12\ 877$).

Mots-clés. Courbe ROC, modèles de prédiction, risques concurrents, multiplicité des tests, test d'hypothèse, validation croisée.

Abstract. Boosted by the growing interest in personalized medicine, the demand for new prediction tools is currently strongly increasing. Many statistical models and strategies have already been discussed to build prognostic tools. Meanwhile, an increasing number of risk factors and new biomarkers are nowadays available for making prediction. In practice, challenge is to choose among different strategies for building prediction models. Fair comparison of prediction strategies is a challenging task.

For comparing two prediction strategies, a commonly applied technique consists of splitting the data once into two data sets: a "learning sample", used to train the two prediction tools, and a "test sample", used to compare them. Unfortunately, the results often depend on how the data were split. Recently, van de Wiel et al. (2009) proposed a testing approach based on multiple splits of the data. The strengths of the approach include its computational ease and universality, which enable to compare arbitrary prediction strategies. It is also general with respect to the prediction accuracy criterion.

Extensions to right censored data and situations with competing risks are discussed. We further provide new insights regarding the underlying assumptions and type-I error control of the original test. Applications to the prediction of cardiovascular events illustrate the potential of the new approach. The aim is to compare risk prediction strategies based on electrocardiogram records. Large data from a Danish cohort are analyzed ($n = 12,877$).

Keywords. Competing risks, hypothesis testing, cross-validation, multiple testing, prediction models, ROC curve.

1 Background

The use of risk prediction scores based on statistical modeling is common in cardiology. For instance, current guidelines use prediction scores to determine whether the risk of stroke is sufficiently high to merit anticoagulation therapy [3]. Typical risk scores in cardiology estimate the probability that a person experiences a cardiovascular event before a time point t (e.g. $t = 5$ years), which is often called the t-year *absolute risk* of cardiovascular events.

Numerous statistical methods have been proposed for building prediction scores based on statistical modeling [10]. Meanwhile, an increasing number risks factors and new biomarkers have emerged. Therefore, there is nowadays room for investigating many different strategies for building prediction scores. From a practical point of view, it remains, however, challenging to compare them.

In this talk, we address the question of testing whether one approach is better than another for building a risk prediction score with a given data set. We consider the general setting in which the two approaches can be completely unrelated to each other. To compare the performance of the two strategies, we focus in this talk on the area under the ROC curve. Our approach enables to properly compare t-year predictions of cardiovascular events. The method handles censored data and properly deals with the competing risk of non-cardiovascular death.

2 Method

Let X be a vector of covariates. As in Gerds and van de Wiel (2011), we define a *prediction strategy* as follows. Based on the training data set $\mathcal{L}_m$, a prediction strategy S_t selects a prediction model $S_t(\mathcal{L}_m)$, such that for every value x, the model predicts the conditional probability that the event of interest occurs before time t given $X = x$. Without loss of generality, at this stage we do not assume any restriction on the prediction strategies S_t^A and S_t^B that we aim to compare.

Hereafter $\theta(t)$ denotes the area under the time-dependent ROC curve at time t, which is a well established concordance index. In our application the interpretation of the value of $\theta(t)$ is the following: *"The probability that the predicted risk of a person who dies from cardiovascular events before time point t is greater than that of a person either alive or dead from non cardiovascular death at time t."*.

2.1 Single split approach

The single-split approach consists of randomly splitting the available data $\mathcal{D}_n$, of size n, into a *learning sample* $\mathcal{L}_m$, of size m, and a *test sample* $\mathcal{T}_{n-m}$, of size $n - m$, i.e,

$$\mathcal{D}_n = \mathcal{L}_m \cup \mathcal{T}_{n-m} \quad \text{with} \quad \mathcal{L}_m \cap \mathcal{T}_{n-m} = \emptyset,$$

where $m < n$. From this partition, the idea consists of training the two rival prediction strategies using $\mathcal{L}_m$ before comparing them using $\mathcal{T}_{n-m}$.

We further denote by $\theta^A_{\mathcal{L}_m}(t)$ and $\theta^B_{\mathcal{L}_m}(t)$ the prediction performances of prediction strategies S_t^A and S_t^B trained on the learning sample $\mathcal{L}_m$. These parameters evaluate the expected performances of the predictions from models $S_t^A(\mathcal{L}_m)$ and $S_t^B(\mathcal{L}_m)$, when applied to an *independent* population of subjects that could benefit from the predictions.

2.2 The single-split test

Using data $\mathcal{T}_{n-m}$, the idea consist of performing a one-sided test for the difference in prediction performance between models $S_t^A(\mathcal{L}_m)$ and $S_t^B(\mathcal{L}_m)$. A one-sided test is usually more relevant than a two-sided because of the asymmetric preference between the two strategies. The corresponding null hypothesis is,

$$\mathcal{H}_0^{\mathcal{L}_m}(t): \qquad \theta^A_{\mathcal{L}_m}(t) \leq \theta^B_{\mathcal{L}_m}(t). \tag{1}$$

Tests for comparing areas under the ROC curve, given a *learning sample*, have previously been suggested. The test of DeLong et al. (1988) covers most of the usual settings, with uncensored data. Another method can be used with survival data, to deal with censoring and competing risks [2]. The latter is used in this talk as we work with right censored survival data.

2.3 Towards the multi-split test

Most of the time, the null hypothesis that we would like to consider is not exactly the one displayed at (1). Instead, this is:

$$\mathcal{H}_0^{\mathcal{D}_n}(t): \qquad \theta^A_{\mathcal{D}_n}(t) \leq \theta^B_{\mathcal{D}_n}(t), \tag{2}$$

which corresponds to the situation where the two prediction strategies are trained on the entire data $\mathcal{D}_n$. Indeed, we want to build prediction models which are as accurate as possible and so we aim to use the entire available data to fit them. The two null hypotheses $\mathcal{H}_0^{\mathcal{L}_m}(t)$ and $\mathcal{H}_0^{\mathcal{D}_n}(t)$ are different because of $\mathcal{L}_m \subsetneq \mathcal{D}_n$. However, they address the same clinical research question. Loosely speaking, they both aim to answer the question: "*Is prediction strategy A better than B for building a clinical prediction model using my data?*".

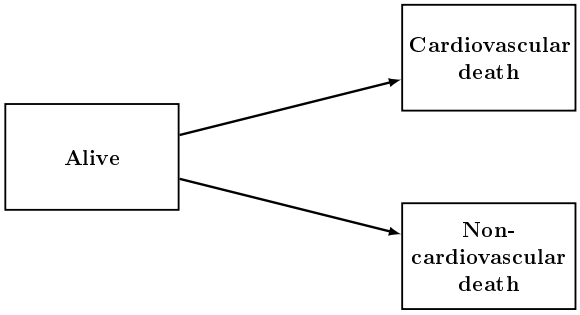
Hereafter, we consider that testing the null hypotheses $\mathcal{H}_0^{\mathcal{L}_m}(t)$ and $\mathcal{H}_0^{\mathcal{D}_n}(t)$ displayed at (1) and (2) is similar enough to be considered as "almost equivalent". Without entering into details, we therefore assume that either **(i)** for the clinical interpretation it does not hurt to see $\mathcal{H}_0^{\mathcal{L}_m}(t)$ as a reasonable approximation of $\mathcal{H}_0^{\mathcal{D}_n}(t)$, for any $\mathcal{L}_m$, or **(ii)** the type-I errors of the tests associated with $\mathcal{H}_0^{\mathcal{L}_m}(t)$ and $\mathcal{H}_0^{\mathcal{D}_n}(t)$ are close enough such as the difference can be neglected. We will get back to this point in the discussion section of our talk.

2.4 Multi-split approach

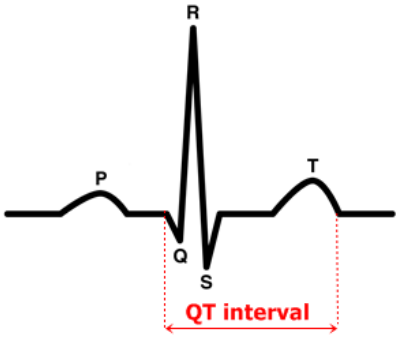
The multi-split approach consists of repeating the single-split approach many times, say $I = 400$ times, before aggregating the results to conclude. It corresponds to the following algorithm:

1. For $i = 1, \ldots, I$:
   - 1.a. Randomly split the data $\mathcal{D}_n$ into $\mathcal{D}_n = \mathcal{L}_m^i \cup \mathcal{T}_{n-m}^i$ with $\mathcal{L}_m^i \cap \mathcal{T}_{n-m}^i = \emptyset$, as in Section 2.1.
   - 1.b. Apply strategies A and B on data $\mathcal{L}_m^i$ to train the prediction models $S_t^A(\mathcal{L}_m^i)$ and $S_t^B(\mathcal{L}_m^i)$, as previously defined.
   - 1.c. Using data $\mathcal{T}_{n-m}^i$, compute p_i, that is the p-value corresponding to the null hypothesis $\mathcal{H}_0^i(t) := \mathcal{H}_0^{\mathcal{L}_m^i}(t): \quad \theta^A_{\mathcal{L}_m^i}(t) \leq \theta^B_{\mathcal{L}_m^i}(t)$, as in Section 2.2.
2. Aggregate $p_1, \ldots, p_I$ and conclude using one of the mathematical results described in the talk. One option, which is based on the following Lemma 1, is: if the median of $\{2p_1, \ldots, 2p_I\}$ is smaller than α, then reject $\mathcal{H}_0(t) := \cap_{i=1}^I \mathcal{H}_0^i(t) \approx \mathcal{H}_0^{\mathcal{D}_n}(t)$ with confidence level α.

Lemma 1. *Let $\gamma \in (0,1)$ and $\widetilde{p}_\gamma = \min\{1, q_\gamma(p_1/\gamma, \ldots, p_I/\gamma)\}$ where $q_\gamma(\cdot)$ is the empirical γ-quantile function and $p_1, \ldots, p_I$ denote I p-values under the null, i.e. $\forall (i,u) \in \{1, \ldots, I\} \times (0,1) \quad \mathbb{P}(p_i \leq u) \leq u$. Then, $\forall \alpha \in (0,1) \quad \mathbb{P}(\widetilde{p}_\gamma \leq \alpha) \leq \alpha$.*



(1) Competing risks setting.



(2) QT-interval on the electrocardiogram.

Proof. A proof is easily derived using arguments as in Meinshausen et al. (2009). □

Remark 1. *The Lemma 1 does not make any assumption on the correlation structure of the p-values. Under this assumption of arbitrary dependencies, it can be shown that the control of the type-I error is sharp [6,9]. Sharper results are available under stronger assumptions [9,11].*

3 Application

The QT interval on the surface electrocardiogram (ECG) represents the time from the beginning of ventricular depolarization to the end of ventricular repolarization (Figure 2). Long QT intervals have been shown to be associated with higher risks of cardiovascular events [8].

We present a simple but pedagogical comparison of two prediction strategies for discriminating subjects at high risk of cardiovascular events. The first is based on the value of the QT-interval only. The second combines the value of the QT-interval and the age of the patient through a regression model. A nonlinear effect of the QT-interval is considered. Separate cause specific regressions for cardiovascular and non-cardiovascular deaths are combined into an absolute 5-year risk prediction model. This modeling strategy is suitable to account for the competing risk setting (Figure 1).

The Data are comprised of $n = 12,877$ men, aged 70-80 years, who had an ECG taken by a general practitioner in the region of Copenhagen [8]. The follow-up covers a period of 11 years (2001-2011). Within the 5 years following the ECG, 23% of subjects were lost to follow-up (censored), 6% died from non-cardiovascular deaths and 15% from cardiovascular events.

The application aims to illustrate and carefully explain the details of the method.

Bibliographie

[1] Andersen, P. K., & Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, 31(11-12), 1074-1088.
[2] Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397
[3] Camm et al, A. (2010). Guidelines for the management of atrial fibrillation: The task force for the management of atrial fibrillation of the european society of cardiology (ESC). *European heart journal*, 31:2369–2429.
[4] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the Areas Under Two or More correlated Receiver Operating Characteristic Curves : A Nonparametric Approach. *Biometrics*, 44(3):837–845.
[5] Gerds, T. A. and van de Wiel, M. A. (2011). Confidence scores for prediction models. *Biometrical Journal*, 53(2):259–274.
[6] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138–1154.
[7] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488).
[8] Nielsen, J. B., Graff, C., Rasmussen, P. V., Pietersen, A., Lind, B., Olesen, M. S., Struijk, J. J., Haunsø, S., Svendsen, J. H., Kø ber, L., Gerds, T. a., and Holst, A. G. (2014). Risk prediction of cardiovascular death based on the QTc interval: evaluating age and gender differences in a large primary care population. *European heart journal*.
[9] Roquain, E. (2010). Type I error rate control for testing many hypotheses: a survey with proofs. *Journal de la Societe Francaise de Statistique*, 152(2):3–38.
[10] Steyerberg, E. (2009). *Clinical prediction models: a practical approach to development, validation, and updating*. Springer.
[11] van de Wiel, M. A., Berkhof, J., & van Wieringen, W. N. (2009). Testing the prediction error difference between 2 predictors. *Biostatistics*, 10(3), 550–560.