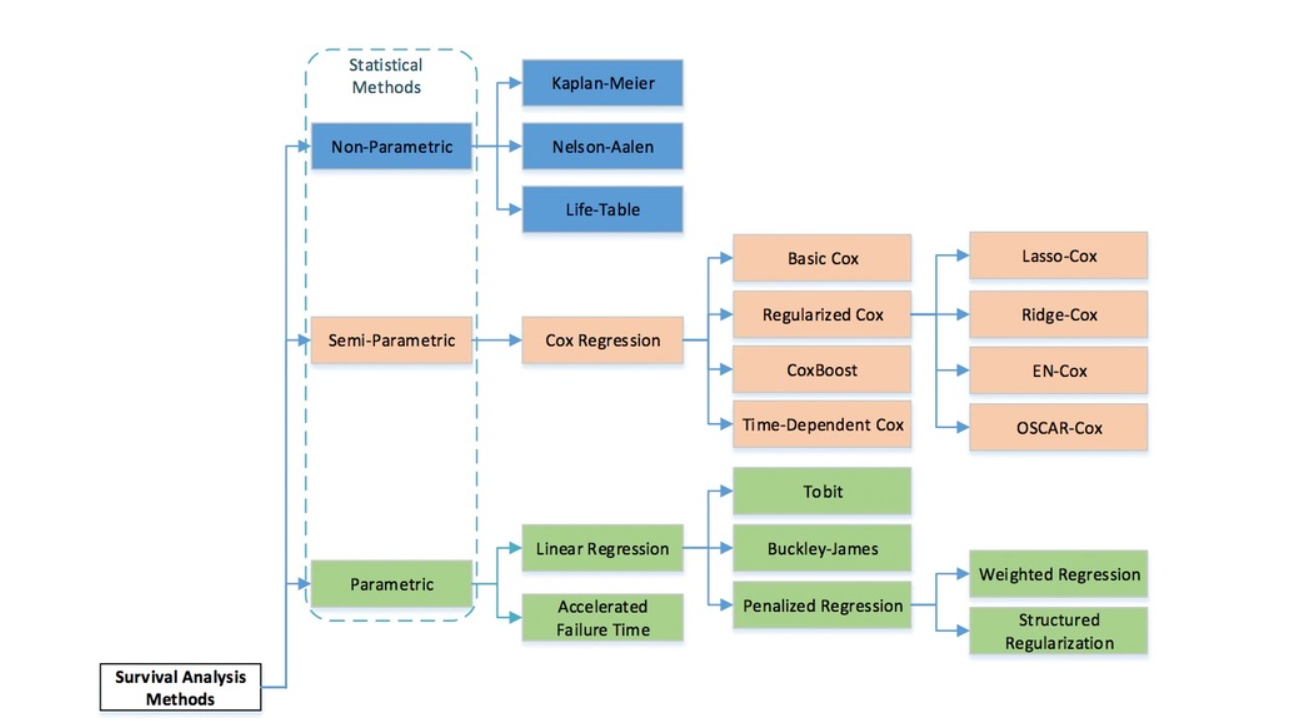
# 深度学习的生存分析文献综述

## 生存分析的常规使用方法：

标准生存分析研究常常分为三部分即非参数，半参数，无参数，其中参数方法依赖于生存时间分布存在严格的概率分布，模型中通过极大似然估计等可进行估计。非参数部分不依赖于潜在分布，生存概率作为关于时间的函数并基于个人总体给定平均视图。非参数的方式中最常见的为**Kaplan-Meier estimator。半参数估计中常常使用 Cox regression model。**



## Kaplan - Meier estimator

1. break the estimation of the survival function S(t)S(t) into smaller steps depending on the observed event times.

## CoxPH

1. 涉及时间以及审查特征，同时涉及协变量的附加数据。（包含所有变量）
2. 半参数，1）协变量（参数部分）2）non-parametric component λ0(t)

CoxPH的基础组成部分是比例假设，即用户死亡次数相互独立，对于每个人的危险曲线不会存在交叉现象，协变量对危险函数存在线性乘法效应

## 时变Cox回归

1. 通过引入时变协变量，即将个人的时间拓展至不同的区间。
2. 假设预测的危害显着依赖于协变量的较晚值而不是基线协变量的值时，应使用时变特征。时变协变量的挑战是在不同时间步长的协变量中缺少数据

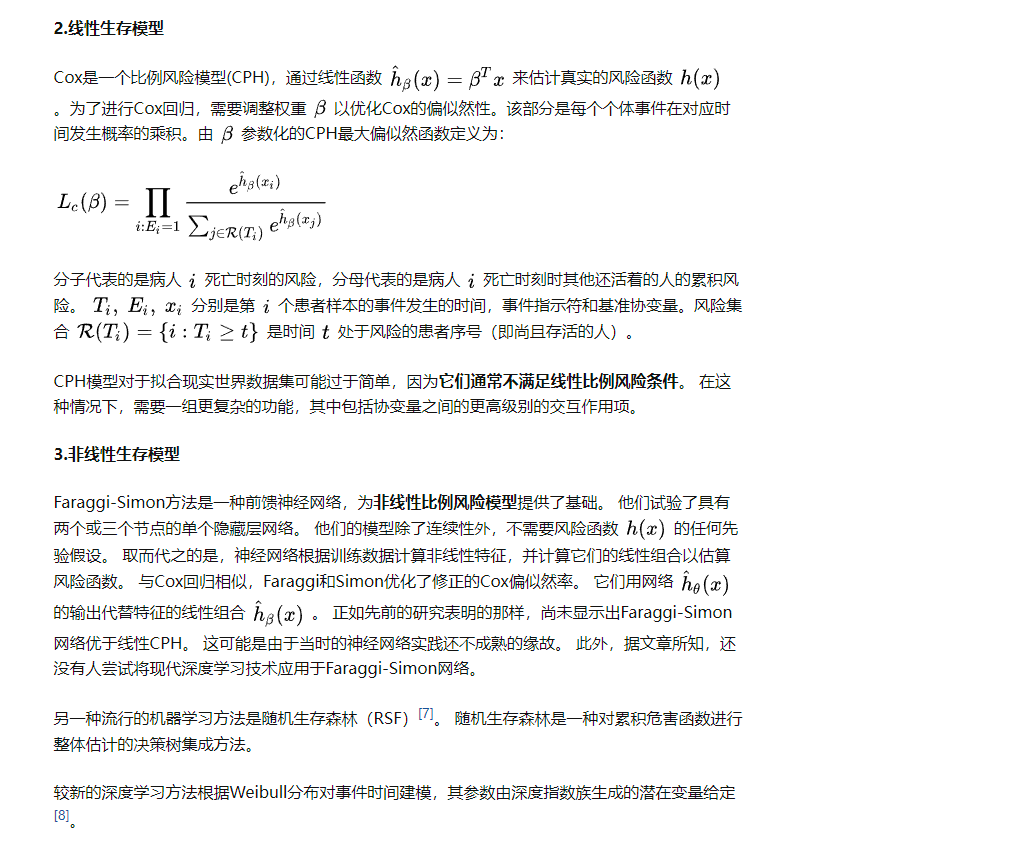
## 随机生存森林

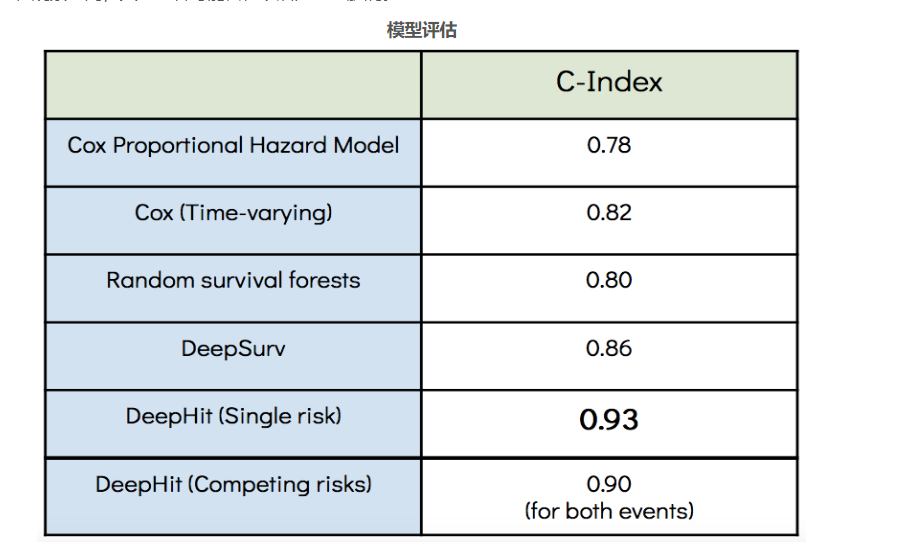
1. 随机生存森林被定义为一种构建累积风险函数的集成估计的树方法。从基础学习器（例如树）构建集成可以显着提高预测性能。

## 六： 深度学习下的生存分析介绍

* 第一个是基于基线 Cox 比例风险模型的进一步发展：**DeepSurv**（第 5.1 节）、**Cox-nnet**（CoxPH 在特定遗传学数据集和正则化上的扩展）。[16]
* 作为一种替代方法，使用 RNN 顺序预测到下一个事件的时间分布的全参数生存模型： **RNN-SURV**、 **Weibull Time-To-Event RNN**等 [17] [26]
* 另一方面，有一些新的高级深度学习神经网络，例如**DeepHit**，也被开发用于处理具有竞争风险的生存数据（第 5.2 节）。

（1）在基于神经网络的生存分析领域存在三种方法，分类，时间编码以及Faraggi-Simon网络的变体。生存数据：基本的协变量x，事件时间与事件指示

* DeepSurv：1. The initial adaptation of survival analysis to meet neural networks (Farragi and Simon, 1995) was based on generalization of the Cox proportional hazards model with only a single hidden layer. 初始模型的主要重点是了解主要协变量与相应的危险风险函数之间的关系。随着 Cox 回归的神经网络架构的发展证明，在现实世界的大型数据集中，变量之间存在非线性交互，保持 Cox 回归模型的主要比例假设是相当复杂的。然而，Farragi 和 Simon 的网络扩展了这种非线性特性。
* DeepHit：通过直接学习到关于存活时间分布的深度神经网络并且对整个潜在随机过程不做任何假设。共享子网与一系列的特定原因的子网。
* 针对两种竞争风险的DeepHit，DeepHit 模型的架构类似于传统的硬参数共享的多任务学习架构，但有两个主要区别。DeepHit 在原始协变量和特定原因子网络的输入之间提供了残差连接。这意味着特定原因子网络的输入不仅是先前共享子网络的输出，而且是原始协变量。这些额外的输入允许特定原因的子网络更好地学习多个原因的非常见表示。
* 



# 生存分析研究报告

生存分析广泛应用于生物信息学与疾病突发的领域，在金融行业以及工程实践性分析上也存在大量的应用。作为统计学的分支，其用于完成对事件发生与持续的时间的预测，例如贷款欺诈，保险流失。在经济学上称为“持续时间分析“，在工程中称作“可靠性分析”。

生存分析的发展可以追溯至17世纪，英国统计学John Grant在1662年制作了第一张生命表。“生存分析”该名词也来源于在生命科学领域中的应用。十九世纪Gompertz和Makeham[1]提出依据时间变化风险率的危险函数分析，生存分析的主要目的是在生存数据中完成对生存函数/风险函数的估计与分析·。二十世纪六七十年代，生存分析主要发展方向为参数化模型，利用极大似然估计方式完成参数估计，但变量的非线性参数难以估计。伴随二十世纪七十年代医学领域的临床试验的出现，过往的参数方法不能满足实践需求，非参数方法中Kaplan-Meier method在生存分析中大放异彩，通过计算在各个时段的生存概率。Cox在1972年提出了比例风险模型（Proportional Hazard Function），CoxPH不仅仅包含了时间以及审核特征同时包含了作为协变量的额外数据，其描述协变量与生存分布之间的关系。因为生存时间资料的分布往往不服从正态分布（大多为正偏态分布），有时甚至不知道它的分布类型，这样就不能采用多重线性回归方法分析。本节介绍的Cox比例风险回归模型（Cox proportional hazard model）可以分析多个因素对生存时间的影响，而且允许有截尾数据存在，是生存分析中最重要的多因素分析方法。不同特征的人群在不同时刻的风险率函数不同，通常将风险率函数表达为基准风险率函数与相应协变量函数的乘积。

基于机器学习的生存分析可以追溯至随机森林模型，随机生存森林被定义为一种构建累积风险函数的集成估计的树方法。从基础学习器（例如树）构建集成可以显着提高预测性能。

**[1] R. H. Wiesner, S. V. McDiarmid, P. S. Kamath, et al. MELD and PELD: application of survival models to liver allocation[J]. Liver transplantation, 2001, 7(7): 567-580**

**[2] D. R. Cox. The analysis of multivariate binary data[J]. Applied statistics, 1972: 113–120**