基于梯度提升树的生存分析优化方法研究及应用

硕士学位论文答辩

答辩人：刘沛（201721060103）

答辩日期：2020 年 5 月 18 日

电子科技大学计算机科学与工程学院

论文的主要工作成果

- 第一作者 | JCR 二区 | 已发表. Hitboost: Survival analysis via a multioutput gradient boosting decision tree method[J].**IEEE Access**, 2019, 7(): 56785-56795.
- 第一作者 | JCR 二区 | 已录用. Optimizing survival analysis of xgboost for ties to predict disease progression of breast cancer[J]. **IEEE Transactions on Biomedical Engineering**, 2020, ():1-1.
- 第二作者 | JCR 二区 | 已发表. Predicting invasive diseasefree survival for early stage breast cancer patients using followup clinical data[J]. **IEEE Transactions on Biomedical Engineering**, 2019, 66(7): 2053-2064.

论文的主要工作成果

其他成果：

- **, 刘沛,**,**,**. 一种基于乳腺癌临床高维数据的分层重要特征选择方法 [P]. 中国, 发明专利,CN201810552686.3,2018 年 5 月 31 日（申请专利）
- **, 刘沛,**,***,**. 一种基于 Efron 近似优化的生存风险建模方法 [P]. 中国, 发明专利,CN201910315815.1,2019 年 4 月 19 日（申请专利）
- **, 刘沛,**,**,**,***. 一种用于生存风险分析的多输出梯度提升树建模方法 [P]. 中国, 发明专利,CN201910315829.3,2019 年 4 月 19 日（申请专利）

目录

研究背景与意义

研究背景与意义

生存分析（a.k.a. time-to-event analysis）

- **研究内容**：个体在不同观测期发生某个特定事件的概率
- **应用领域**：医疗健康、金融等

在临床疾病预后研究中，生存分析方法用来

- 分析患者随访数据
- 建立生存预后模型
- 辅助诊疗
- 发现疾病重要影响因子

研究背景与意义

生存分析预后模型示例

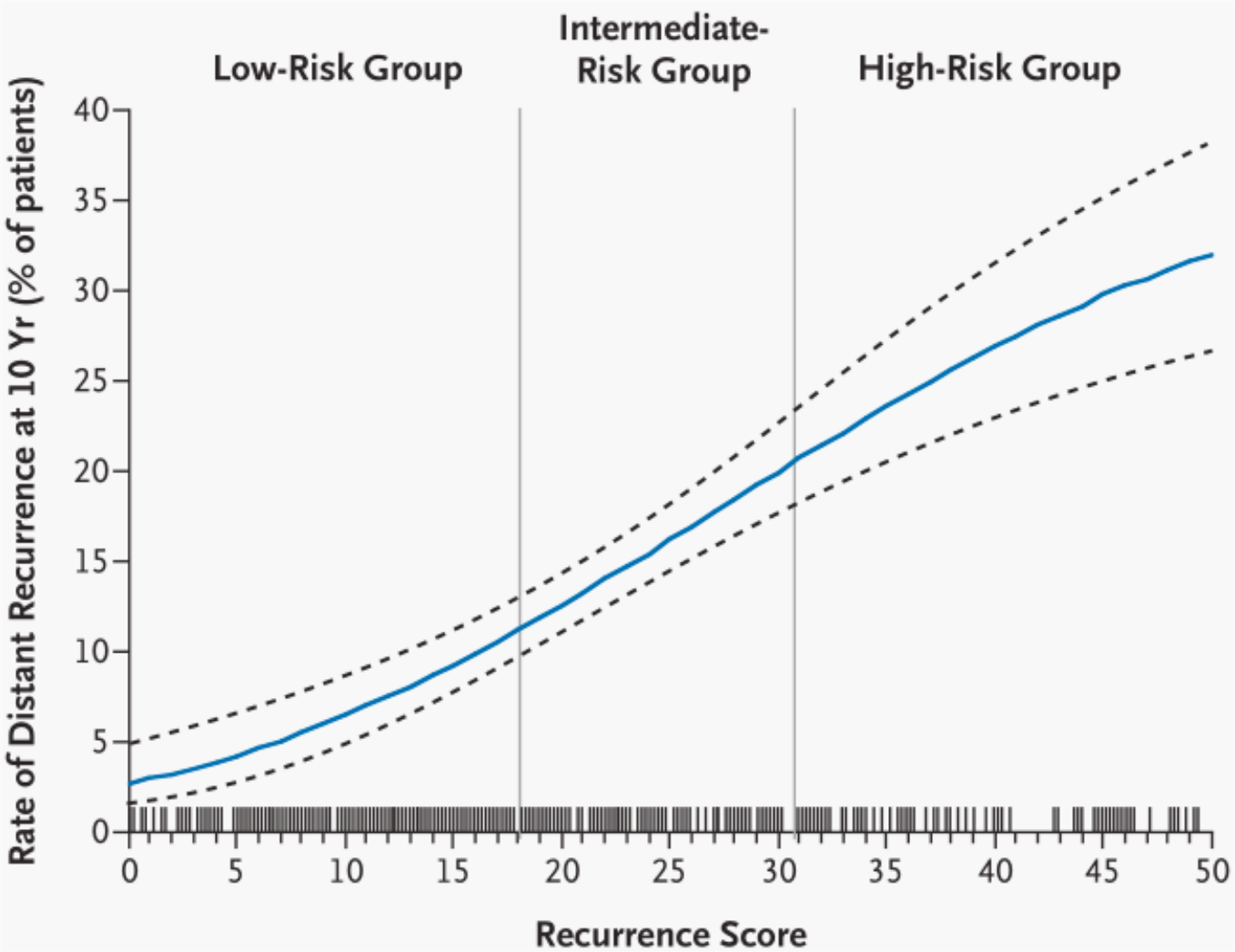

Figure 1: 21Gene - 乳腺癌复发风险评分模型

研究历史与现状

生存分析基础

生存数据 $\{(x_i, T_i, \delta_i) \mid i = 1, \ldots, n\}$

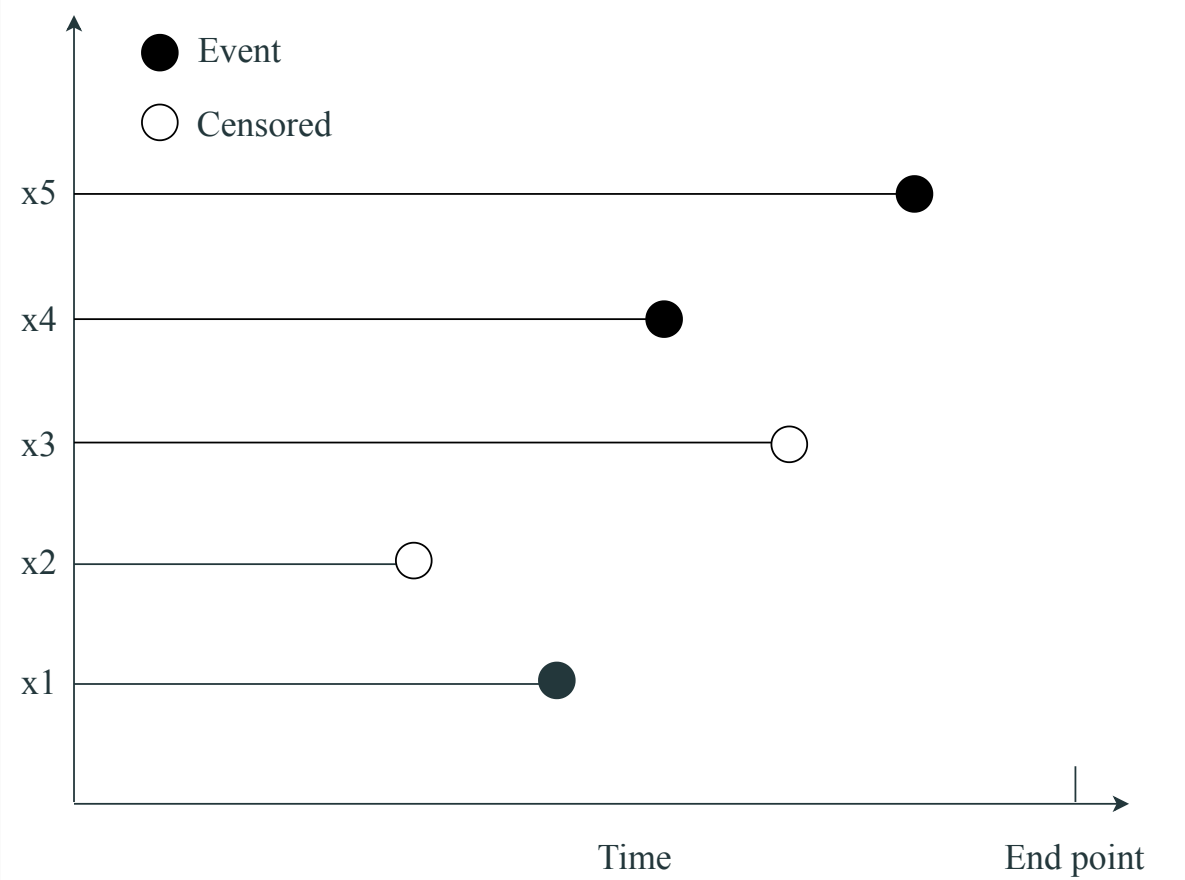

Figure 2: 生存数据示意图

生存分析基础

生存函数 $S(t)$ 与风险函数 $h(t)$：$h(t) = \frac{-S'(t)}{S(t)}$

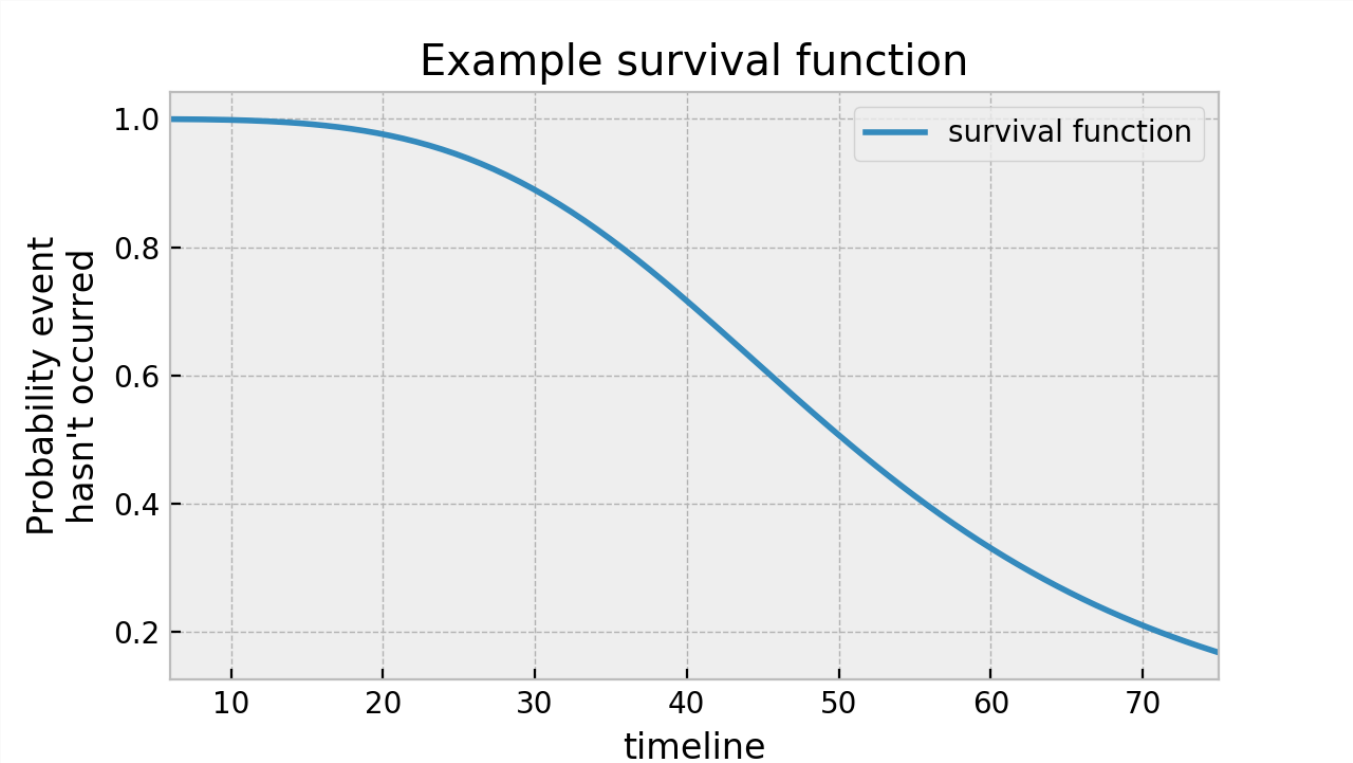

Figure 3: 生存函数曲线示意图

生存分析方法

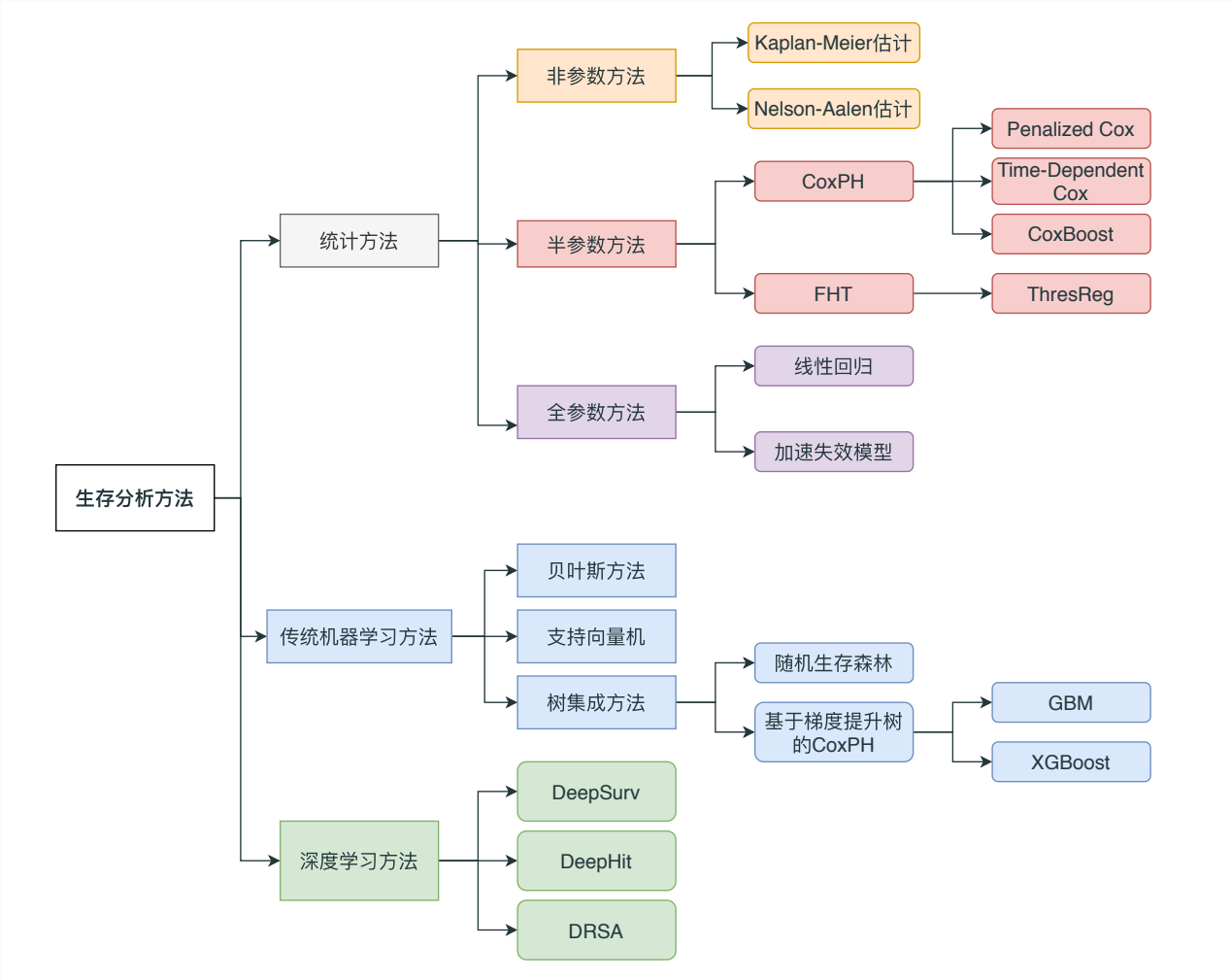

Figure 4: 常见的生存分析方法

论文的主要工作及贡献

主要工作及贡献

论文的主要工作及贡献

- *HitBoost* 生存分析方法
- *BecCox* 生存分析方法
- 乳腺癌复发预后模型

第一部分：HitBoost 生存分析方法

本节内容参见文献：*Pei Liu, **, **. Hitboost: Survival analysis via a multi output gradient boosting decision tree method[J]. IEEE Access, 2019, 7: 56785-56795*。

针对现有生存分析方法依赖先验假设或解释性不足的问题，本文在传统的 FHT 模型的基础上研究并提出了一种基于梯度提升树的生存分析方法：HitBoost。

HitBoost 方法的贡献和创新

- 使用多输出的梯度提升树建模，提高了模型的表达能力
- 引入生存分析中极大似然估计函数和凸函数近似的一致性指数作为联合目标函数，提高了预测性能
- 不再遵循任何先验假设，提升了算法的应用场景
- 仍然具有一定的解释性，保证了模型的实用性

第一部分：HitBoost 生存分析方法

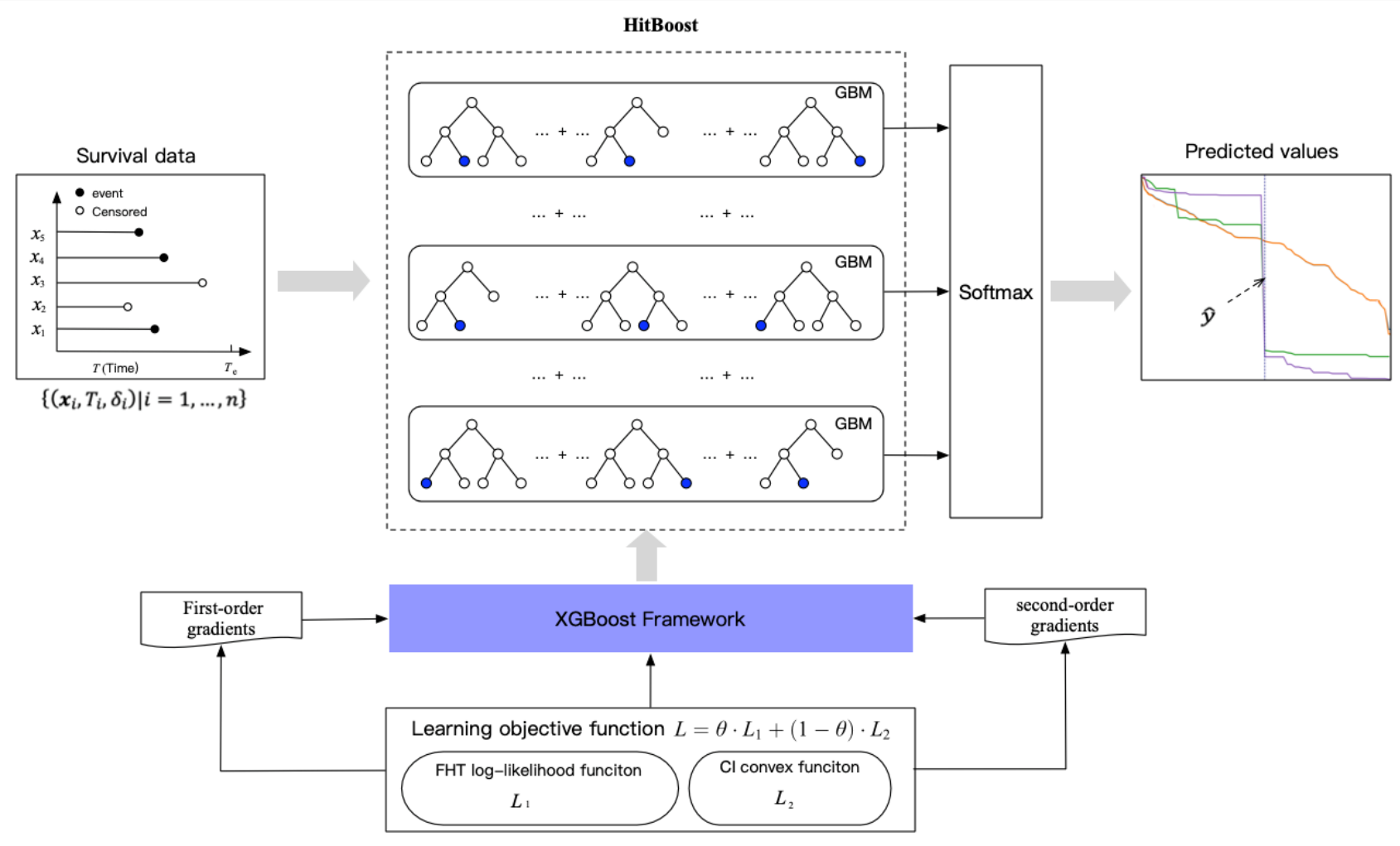

Figure 5: HitBoost 模型框架示意图

第一部分：HitBoost 生存分析方法

HitBoost 模型目标函数 $L = \theta \cdot L_1 + (1-\theta) \cdot L_2$，其中

- L_1，*FHT* 模型中的极大似然估计函数
- L_2，凸函数近似的一致性指数

L_2 中使用如下凸函数近似一致性指数，有效避免模型过拟合。

$$\phi(x,y) = \begin{cases} [-(x-y-\gamma)]^n & \text{if } x-y < \gamma, \\ 0 & \text{if } x-y \geq \gamma. \end{cases}$$

HitBoost 方法实现的主要步骤

- 推导目标函数关于 $\hat{y}$ 的一阶和二阶梯度，见定理 *3.1-3.4* 及证明
- 借助 *XGBoost* 梯度提升树框架现实了 *HitBoost* 方法

第一部分：HitBoost 生存分析方法

实验公开数据集如下，数据预处理及划分流程见论文 3.3 节。

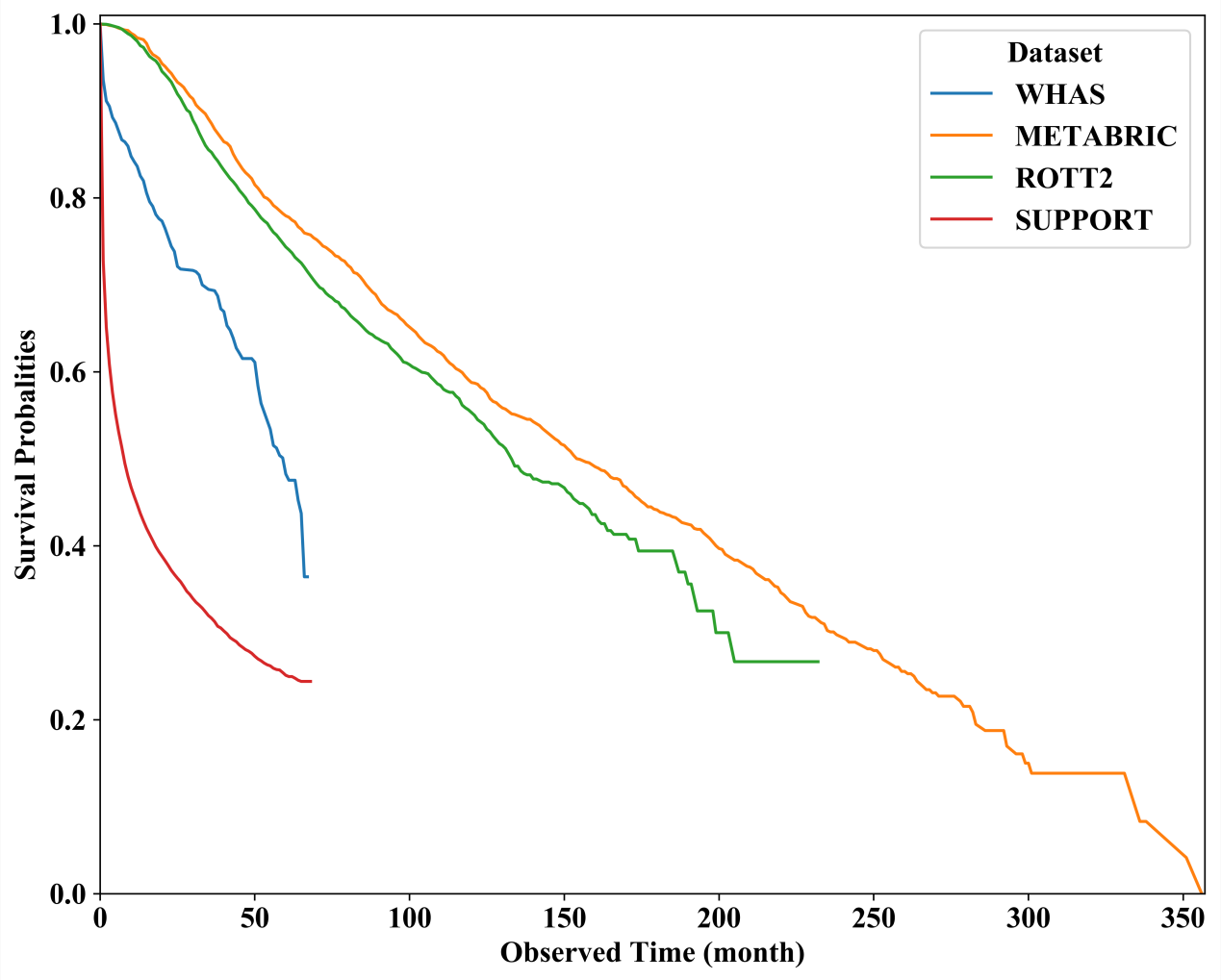

Figure 6: 实验数据集

第一部分：HitBoost 生存分析方法

HitBoost 模型使用的超参数通过贝叶斯超参数优化方法得到，详见论文 3.4.1 节实验设置。

实验结果（一致性指数）

表 3-4 HitBoost 模型性能对比（星号“*”表示我们提出的方法）

方法	WHAS	SUPPORT	METABRIC	ROTT2
CoxPH	0.740648	0.593005	0.633109	0.698081
CoxBoost	0.740682	0.590609	0.624546	0.698542
ThresReg	0.732674	0.591483	0.621219	0.658560
RSF	0.913789	0.614945	0.650566	0.675589
HitBoost*	**0.929190**	**0.631281**	**0.668679**	**0.705427**

Figure 7: HitBoost 模型性能对比

样例分析：模型对个体生存曲线预测示例

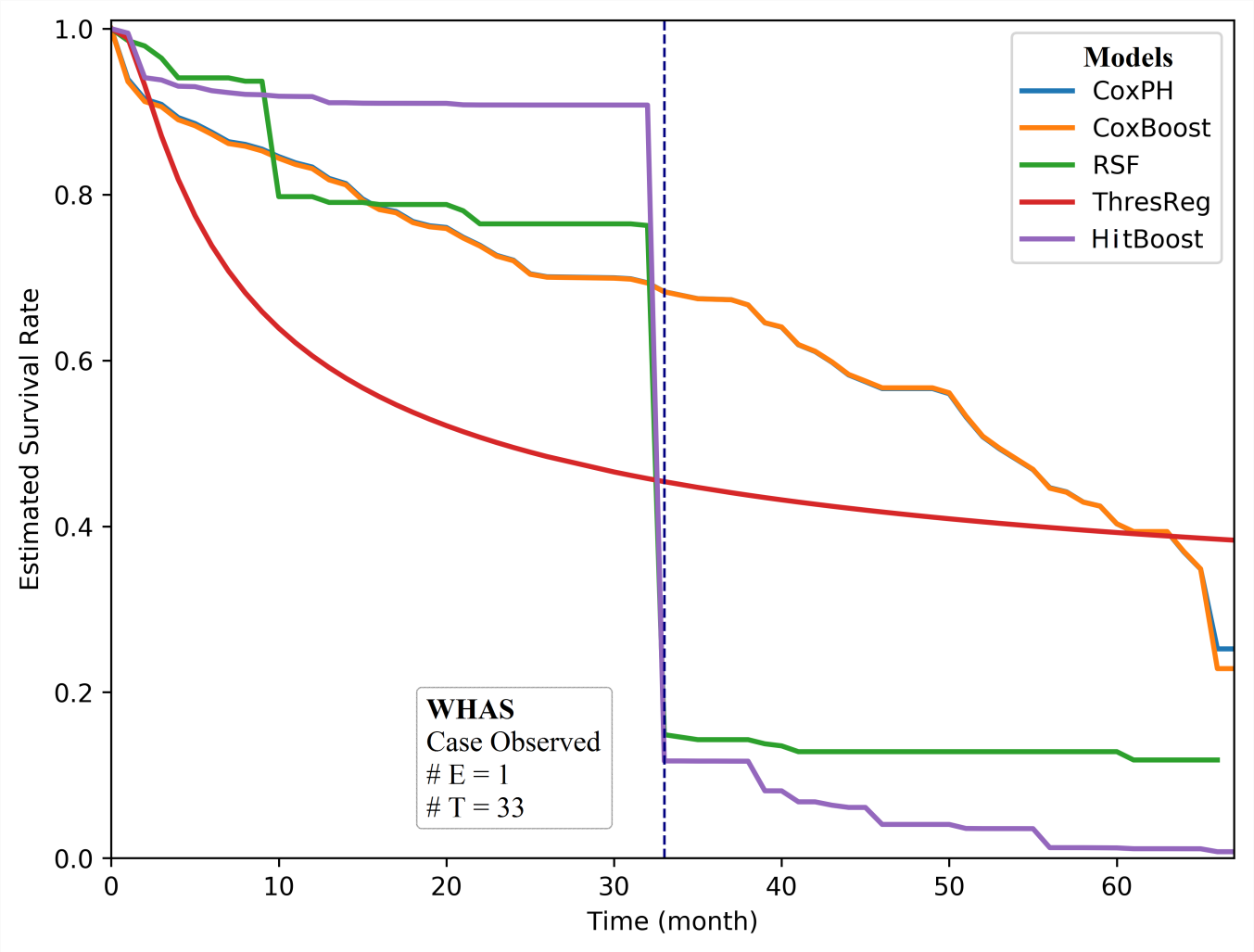

Figure 8: WHAS 数据集

第二部分：BecCox 生存分析方法

考虑到 Cox 流派生存分析方法存在的不足，如偏似然估计函数不够精确以及模型容易过拟合，本文在传统 Cox 比例风险模型基础上研究并提出了基于梯度提升树的优化方法：BecCox。

BecCox 方法的贡献和创新

- 遵循比例风险假设，预测事件发生的风险比例，可广泛应用于传统 *Cox* 生存分析场景
- 在目标函数上，使用更加精确的偏似然估计函数，联合调整风险排序的一致性指数，缩小了目标函数给模型预测带来的偏差
- 在 *Cox* 流派的方法中，相比经典的 *Cox* 比例风险系列模型，该方法有着更好的风险预测性能

第二部分：BecCox 生存分析方法

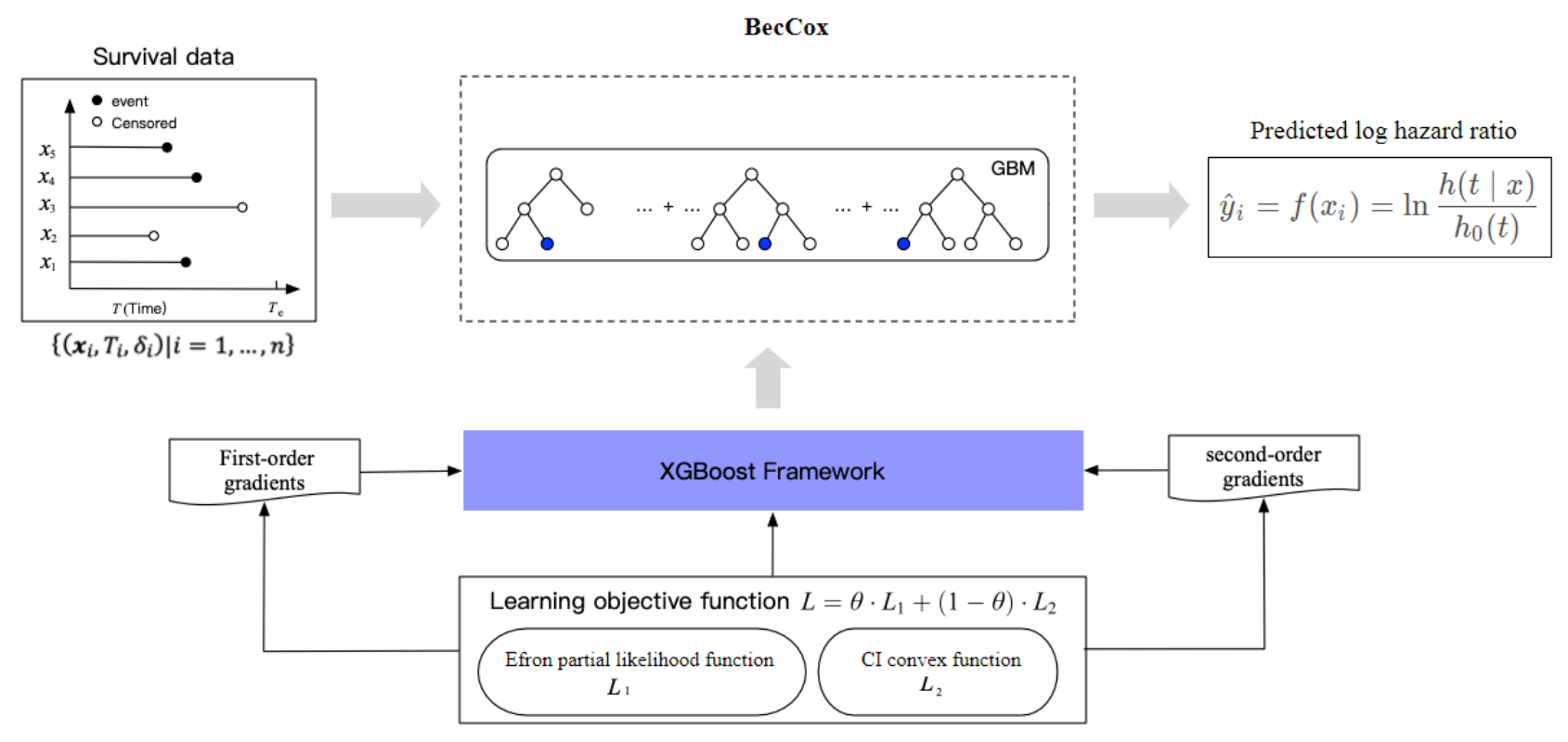

Figure 9: BecCox 模型框架

第二部分：BecCox 生存分析方法

BecCox 模型目标函数 $L = \theta \cdot L_1 + (1 - \theta) \cdot L_2$，其中

- L_1，*Efron* 近似的偏似然估计函数的负对数
- L_2，凸函数近似的一致性指数（同 *HitBoost*）

BecCox 方法实现的主要步骤

- 推导目标函数关于 $\hat{y}$ 的一阶和二阶梯度，见定理 *4.3-4.6* 及证明
- 借助 *XGBoost* 梯度提升树框架现实了 *BecCox* 方法

第二部分：BecCox 生存分析方法

BecCox 模型使用的超参数通过贝叶斯超参数优化方法得到，见本文 4.3.1 节实验设置。

实验结果（一致性指数）

表 4-1 BecCox 模型性能对比（星号“*”表示我们提出的方法）

方法	WHAS	SUPPORT	METABRIC	ROTT2
CoxPH	0.740648	0.593005	0.633109	0.698081
CoxBoost	0.740682	0.590609	0.624546	0.698542
CoxNet	0.740340	0.560036	0.633027	0.698391
GBM	0.894794	0.621311	0.643049	0.687475
BecCox*	**0.898320**	**0.631837**	**0.645986**	**0.702102**

Figure 10: BecCox 模型性能对比

从理论上来看，当数据中出现的 Ties 越多时，BecCox 方法对目标函数的近似越精确，对风险比例的估计偏差越小。

第三部分：乳腺癌复发预后模型

乳腺癌临床数据 WCH

- 四川大学华西医院乳腺疾病临床研究中心乳腺癌信息管理系统
- *5293* 例早期乳腺癌患者记录，*15* 个乳腺癌临床特征

WCH 数据集的处理和特征筛选工作，参见文献：**, Pei Liu, **, et.al. Predicting invasive diseasefree survival for early stage breast cancer patients using followup clinical data[J]. IEEE Transactions on Biomedical Engineering, 2019, 66(7): 2053-2064.

第三部分：乳腺癌复发预后模型

模型性能：使用 HitBoost 方法建立的乳腺癌复发预后模型在 WCH 测试集上的表现为 0.72323，而 CoxPH、CoxBoost、ThresReg、RSF、BecCox 方法在同一测试集上的表现分别为 0.69354、0.68752、0.66792、0.70517、0.71029。

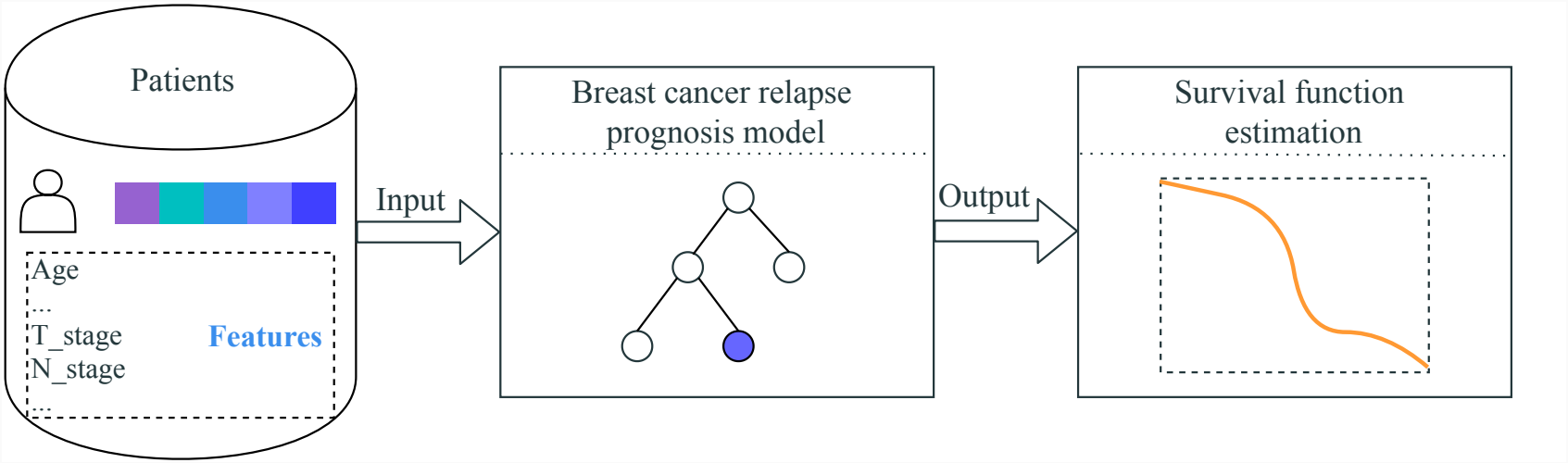

Figure 11: 乳腺癌复发预后模型应用流程

第三部分：乳腺癌复发预后模型

模型应用：探究对早期乳腺癌患者复发有重要影响的因子

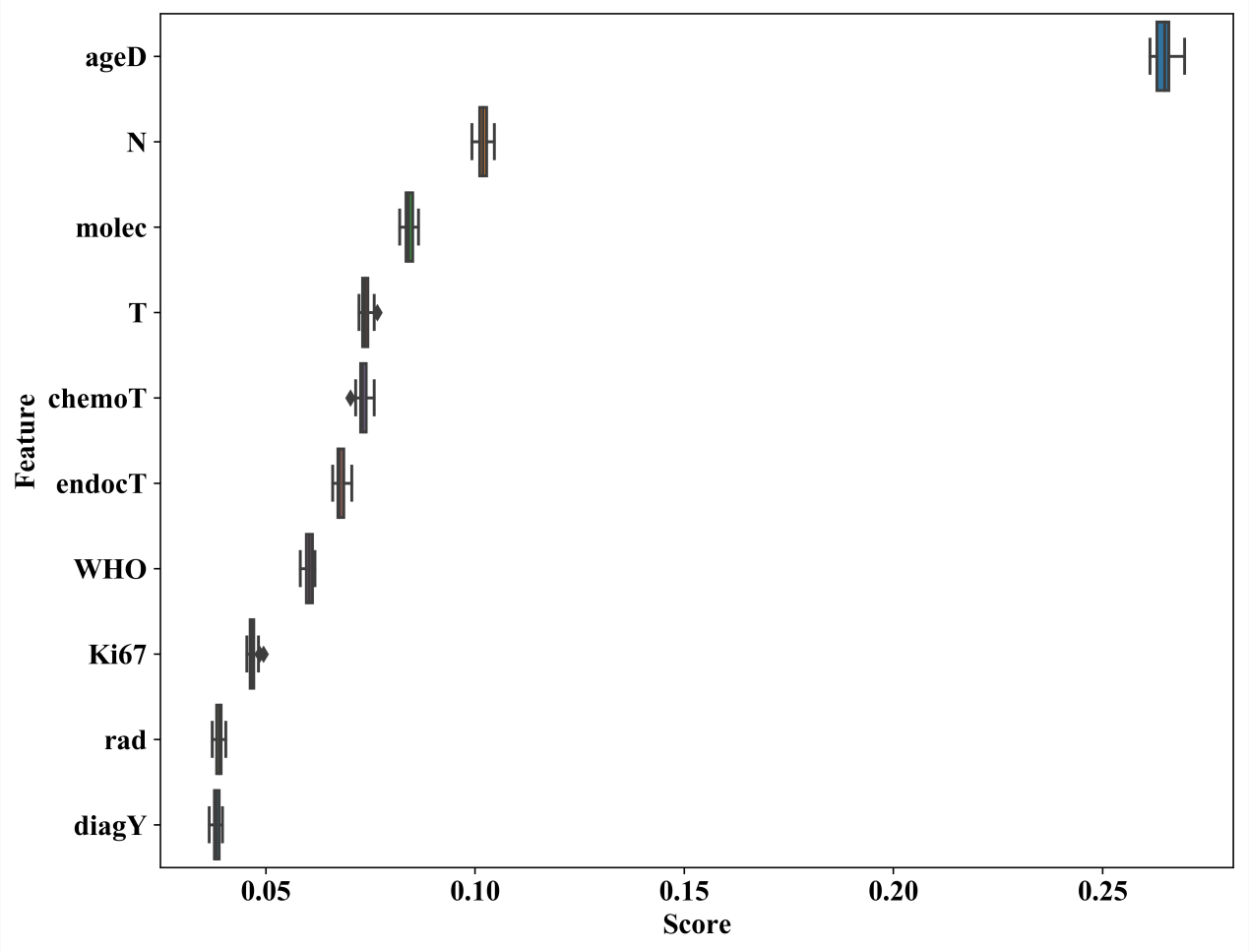

Figure 12: 重要影响因子排序

第三部分：乳腺癌复发预后模型

模型应用：内分泌治疗 + 化疗推荐

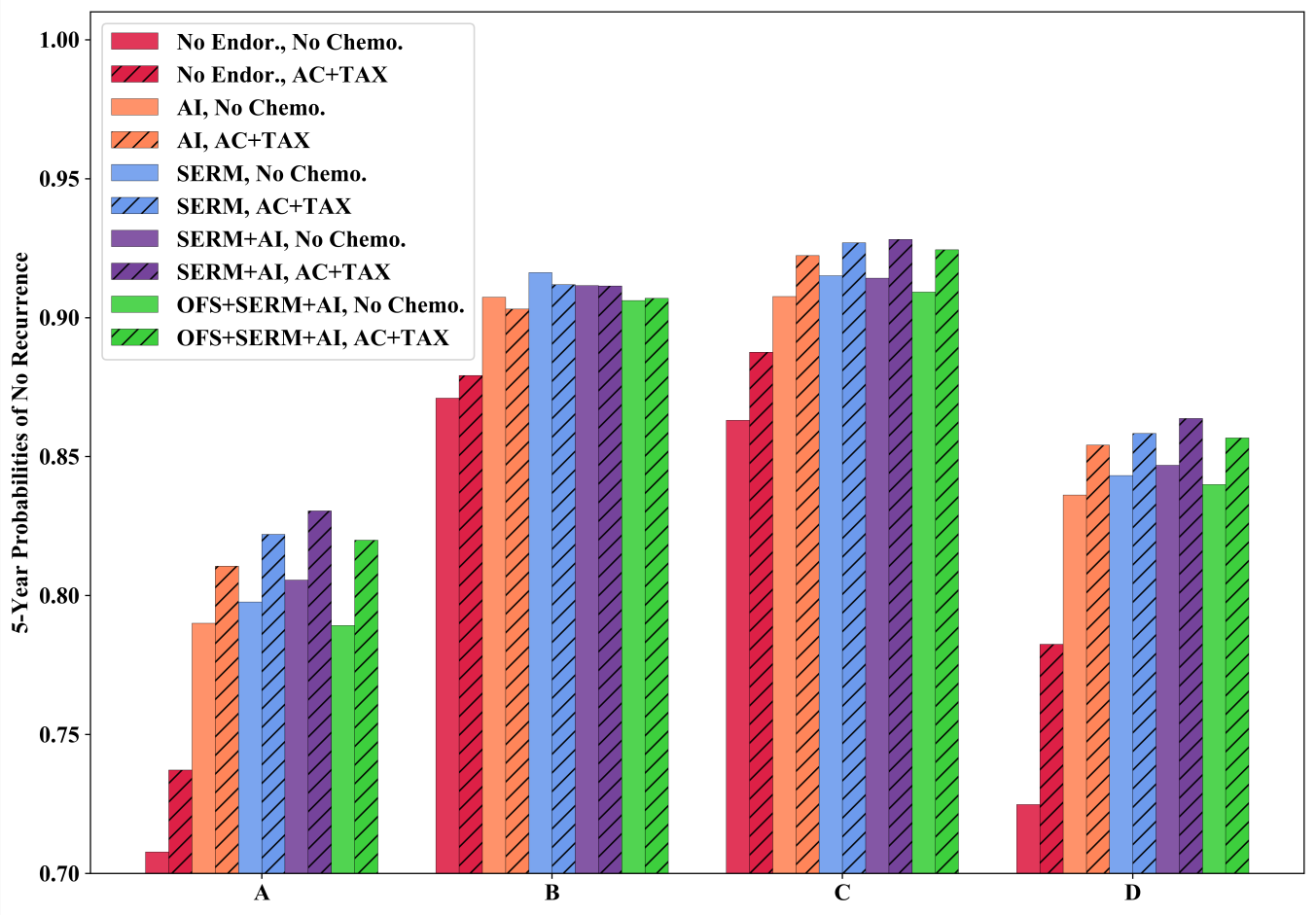

Figure 13: 治疗推荐示例

谢谢各位老师