艾滋病药物疗效影响因素分析

——基于一项双盲安慰剂对照试验数据的生存回归分析

于沁涵　1910227

一、背景介绍

艾滋病是由感染艾滋病病毒（HIV）引起。HIV 是一种能攻击人体免疫系统的病毒，它把人体免疫系统中最重要的 CD4T 淋巴细胞作为主要攻击目标，大量破坏该细胞，使人体丧失免疫功能。因此，人体易于感染各种疾病，并可发生恶性肿瘤，病死率较高。更值得关注的一个问题是 HIV 在人体内的潜伏期平均为 8～9 年，在艾滋病病毒潜伏期内，可以没有任何症状地生活和工作多年。

艾滋病病毒的传播途径有血液传播、性接触传播、母婴传播三条途径，其中占比最多的就是性接触传播。如图 1，2017 年性传播占比 69.57%。据统计表明，天津市平均 10 名患者中就有 7 名因为同性恋性传播，平均每周有一名学生患艾滋病。

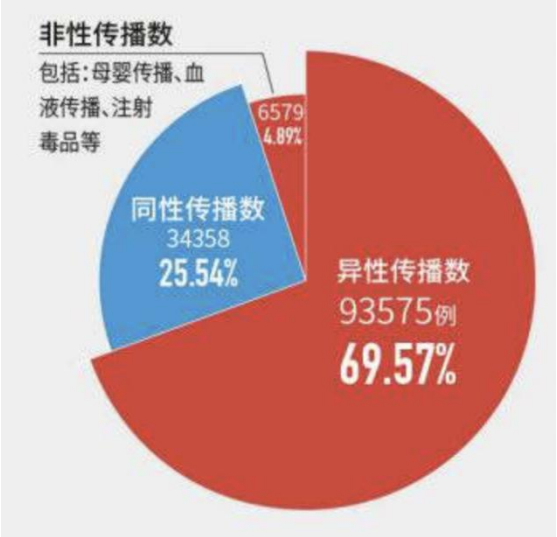

图 1 2017 年艾滋病传播途径占比

目前，艾滋病确诊病例仍然没有得到有效的控制，而因为艾滋病丧失生命的人也不在少数，这令更多医学研究者开始考虑研究出更多可以治疗艾滋病的治疗方案。面对艾滋病的治疗，目前仍没有可以完全治愈艾滋病的治疗方案，但是有可以控制病情恶化的治疗方案。药物治疗是其中最广泛的一种，目前国际上有 6 类 30 多种药物可以治疗艾滋病，分别为核苷酸反转录酶抑制剂(NRTLs)、非核苷酸反转录酶抑制剂(NNRTLs)、蛋白酶抑制剂(PLs)、整合酶抑制剂(INSTLs)、融合酶抑制剂(FLs)及 CCR5 抑制剂，其中最推荐的治疗方案包括两种 NRTLs(TDF+3TC+FTC/TAF)和第三类药物。然而，针对不同的艾滋病患者，患者可能会因为免疫力下降而导致其他疾病感染，而不同性别、种族、患病史也都

可能影响药物治疗的效果.....因此针对药物治疗的方案，仍需要更深一步的探索与研究。

为了探究艾滋病不同药物疗效的影响，我们找到了一个数据来进行分析与研究。数据来自一项双盲安慰剂对照试验，该试验比较了不同药物方案对 HIV 感染患者的疗效。治疗方案可能涉及药物有印地那韦（IDV）、齐多夫定（ZDV）或司他夫定（d4T）、拉米夫定（3TC）等。鉴于实际意义和数据的可获得性，本文基于数据进行生存回归分析。利用生存曲线、Cox 比例风险模型对数据进行分析，探究艾滋病药物疗效影响问题，并得到结论。

二、数据来源和说明

本文的样本数据来自一项双盲安慰剂对照试验，数据比较了不同药物方案对HIV 感染患者的疗效，一个包含 1151 个观测。每个观测代表一个患者，这里面的患者设定每立方毫米的 CD4 细胞数不超过 200 个，且至少接受过 3 个月的齐多夫定治疗。

确诊事件或死亡的时间代表了感染 HIV 患者从实验开始到确诊或者死亡的天数，也是本文的主要研究目的，因此为因变量。自变量包括治疗中止指征、筛查时 CD4 层、性别等 10 个变量，具体变量说明见表 1。

表 1 变量说明

	变量名称	详细说明	取值范围	备注
因变量	确诊时间或死亡天数	连续变量	1～364	艾滋病确诊时间或死亡天数
	死亡天数	连续变量	单位：天	只取整数
自变量	治疗中止指征	定性变量（2 水平）	1 =治疗（包括 IDV） 0 =对照组	治疗组占比 49.8%
	筛查时 CD4 层	定性变量（2 水平）	0 = CD4 <= 50 1= CD4 > 50	CD4 > 50 占比 61.8%
	性别	定性变量（2 水平）	1 =男性 2 =女性	男性占比 82.6%
	种族/民族	定性变量（5 水平）	1 =非西班牙裔白人 2 =非西班牙裔黑人 3=西班牙裔 4=亚裔，太平洋岛民 5=美洲印第安人	非西班牙裔白人占比 51.7% 非西班牙裔黑人占比 28.4% 西班牙裔占比 17.6% 亚裔，太平洋岛民占比 1.2% 美洲印第安人占比 0.9%
	静脉用药史	定性变量（3 水平）	1 =没有 2 =目前 3 =之前	没有静脉用药史占比 84.1% 之前有静脉用药史占 15.5%
	是否患有血友病	定性变量（2 水平）	1=是 0=否	患有血友病占比 3%
	生活质量评分	定性变量（4 水平）	100 =正常 90 =可能正常活动 80 =正常活动和努力 70 =关心自己	正常占比 34.4% 可能正常活动占比 47% 正常活动和努力占比 15.8%
	基准 CD4 细胞计数	连续变量	0～392	
	之前几个月使用 ZDV	连续变量	3～312	只取整数
	年龄	连续变量	15～73	只取整数

三、描述性分析

将数据读入后，我们对数据进行描述性分析，绘制了各变量的生存曲线。利用 R 语言进行编程，得到结果。在这里，我们先做一个假设：假设疾病发展速度指的是由试验开始到确诊或死亡事件的时间，即数据表中的 censor。

（一）因变量：确诊或死亡时间(time)、死亡时间(time_d)

确诊或死亡时间指的是从试验开始到患者确诊死亡、或试验截止的天数。对感染 HIV 患者，我们考察进入临床研究开始至确诊 AIDs 时长。对于数据，一共有完全数据、右删失数据两种数据，利用确诊或死亡时间指示器作为判定标准，绘制生存曲线，得到结果如图 2。由图 2 可以看出患者到确诊的过程存活率很高，也意味着该病发展平均周期较长，在 300 天左右时，仍有 90%的患者未确诊，这与艾滋病潜伏期长相照应。

死亡时间指的则是从试验开始到死亡或试验截止的天数。对感染 HIV 患者，我们接着考察入临床研究开始至死亡时长，绘制生存曲线，得到结果如图 3。由图 3 可以看出患者的存活率非常高，在 300 天的时候，患者存活率仍然可以到 96%以上，其平均存活天数可达 242 天左右。

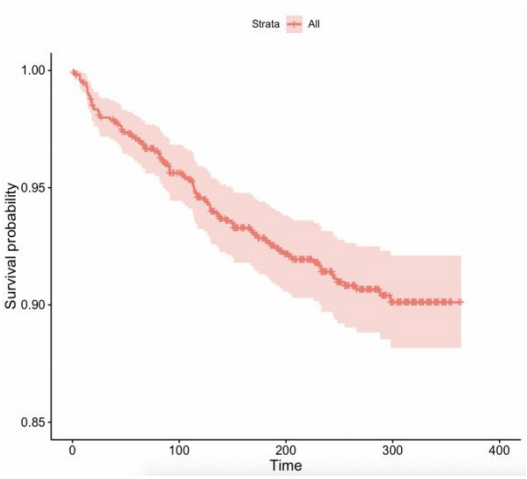

图 2 确诊或死亡时间的生存曲线

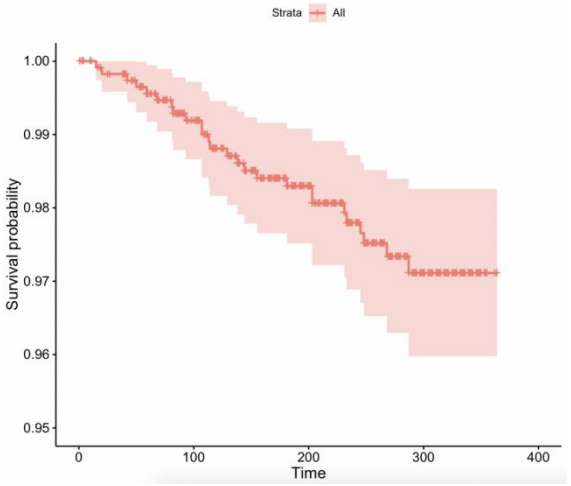

图 3 死亡时间的生存曲线

（二）自变量：性别

我们对性别为男、女的患者进行分组绘制生存曲线，观察其疾病发展速度，得到结果图如下。从图中可以得到结论：二者生存率均呈下降趋势，早期下降较快，后期趋于平稳。在试验初期，男女生存概率基本相同，在 220 天以后，女性的生存概率更大，这说明女性疾病发展速度更慢。总而言之，两者差异没有那么明显，两者的生存率都很高，都在 85%以上。

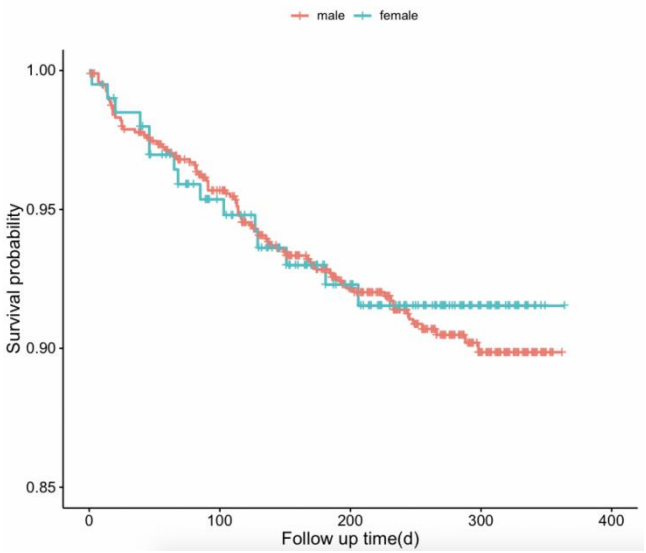

图 4 性别分组生存曲线

（二）自变量：种族

不同的种族也可能对药物治疗有着一定的影响作用。我们将患者分为五个种族非西班牙裔白人、非西班牙裔黑人、西班牙裔、亚裔，太平洋岛民和美洲印第安人，分别进行分组绘制生存曲线，观察其疾病发展速度，得到结果图如图 5。

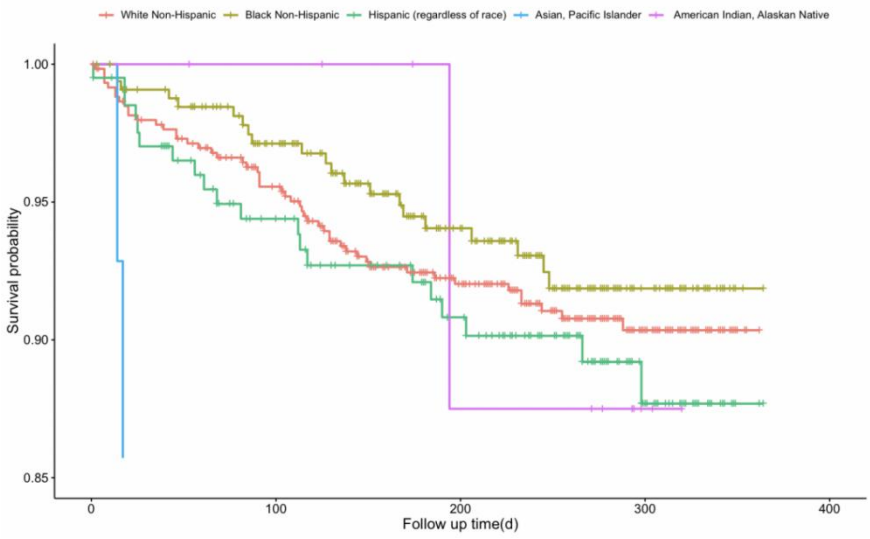

图 5 种族分组生存曲线

从图中可以得到结论：五个种族的生存率均呈下降趋势，不同种族之间生存

概率差异较大。亚裔的生存概率最低，也就是疾病发展速度最快；太平洋岛民和美洲印第安人在 200 天以前疾病发展较其他 4 组更更慢但是 200 天左右迅速恶化。而其他 3 组的生存概率都是开始下降，最后平稳的趋势，其中非西班牙裔黑人疾病发展速度较其他两个更慢，西班牙裔在非西班牙裔白人、非西班牙裔黑人、西班牙裔中生存概率最低，疾病发展速度最快。5 个种族的生存率都很高，都在 85%以上。

（三）自变量：生活质量评分

生活质量评分，又叫卡诺夫斯基表现量表(KPS 得分)。它表示患者对目前症状耐受程度进行自我主观评价的定量评分系统，数据共分 4 个等级。对 4 个等级的得分分别进行分组绘制生存曲线，得到结果如图 6。由图 6 知，随着生活质量评分下降，患者的生存概率也在下降。这意味着 70 分的患者生存概率最低，100 分的患者生存概率更高，这是因为 70 分的患者在心态上更容易崩溃，主观能动性较差，会导致病情恶化，而 100 分的患者没有倾诉没有抱怨，更容易在治疗上取得效果。总而言之，KPS 得分的生存概率均高于 50%。

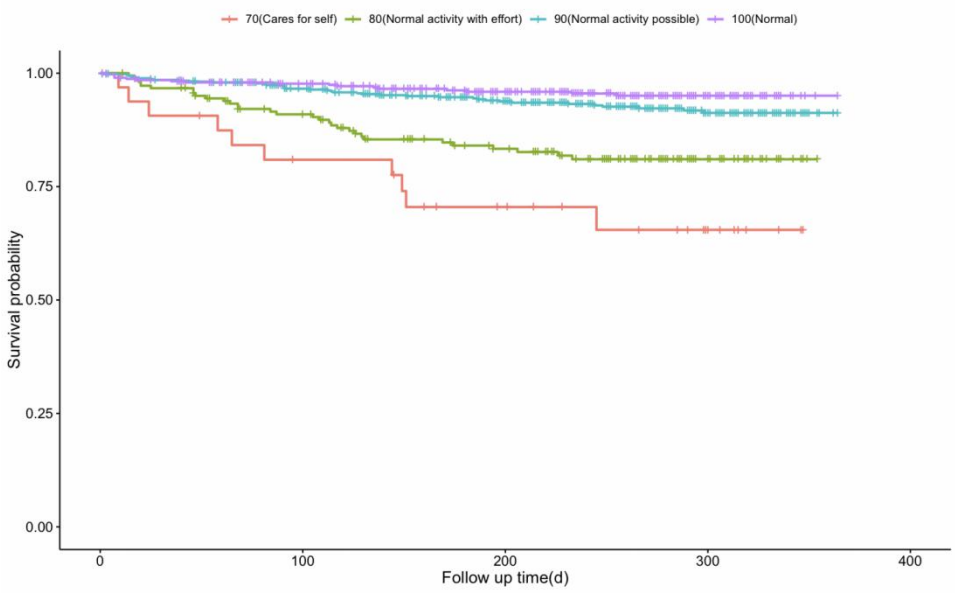

图 6 生活质量评分分组生存曲线

（四）自变量：筛查时 CD4 层

CD4 细胞是人体免疫系统中的一种重要免疫细胞，由于艾滋病病毒攻击对象是 CD4 细胞，因此患者要进行 CD4 检测来查看目前免疫细胞的数量，其检测结果对艾滋病治疗效果的判断和对患者免疫功能的判断有重要作用。在本文中，我们设置 CD4 的一个阈值为 50，超过 50/cm^3 时便认为免疫细胞多，少于 50/cm^3 认为免疫细胞少。绘制生存曲线，结果如图 7。由图 7 看出 CD4 超过 50/cm^3 组的生存概率明显高于少于 50/cm^3 的组，这说明人体内免疫细胞少的时候会加快疾病发展的速度，这与事实相符。

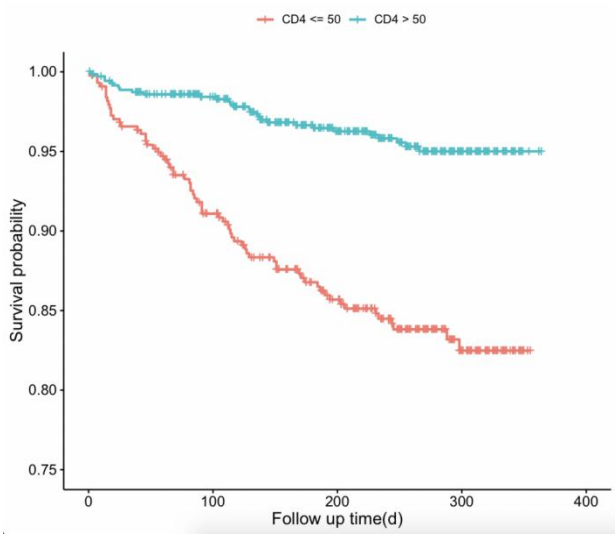

图 7 筛查时CD4 层分组生存曲线

（五）自变量：治疗中止指征

在艾滋病的治疗中，茚地那韦（IDV）作为高效抗逆转录病毒疗法的组成部分，可能对艾滋病的治疗起到一定效果。为了验证其效果，我们对治疗中止指标进行分析，它分为包含 IDV 和对照组两组，我们绘制生存曲线得到如下结果。对于图 8，可以发现对于使用 IDV 的组别，其生存概率更高，说明 IDV 可以起到控制艾滋病病情发展的作用。同样，两者的存活率都在 85%。

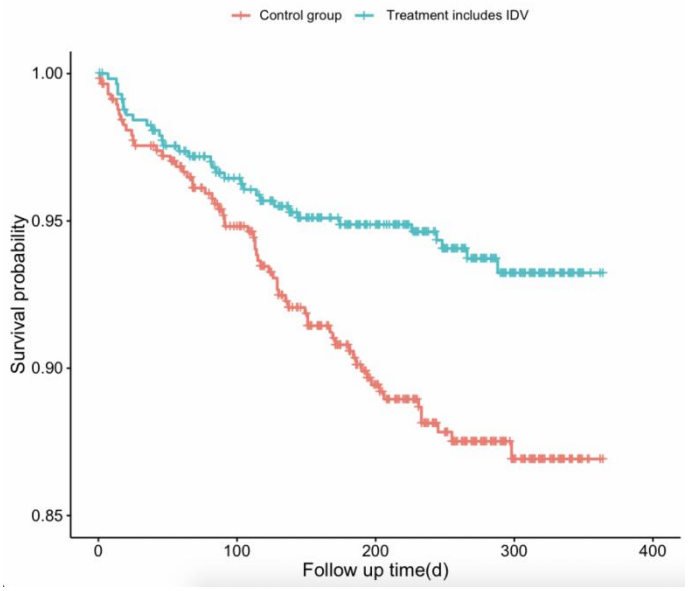

图 8 是否包含IDV分组生存曲线

（六）自变量：静脉用药史

静脉用药是给静脉注射药物的一种治疗方法，考虑到药物之间的相互作业，不难想象有用药史或者正在用药的患者都有可能影响艾滋病药物治疗的效果，因此我们采用静脉用药史这个变量对之前用药史、当下正在使用以及从未使用的患者人群进行分析。

由图 9 知，以前有静脉用药史患者的生存概率比从未有静脉用药史的患者生存概率更高，而目前正在使用的患者在 100 多天的时候突然下降，生存概率最低，说明静脉用药与治疗方案之间的药物可能存在相互影响，且影响效果消极。因此，我们认为静脉用药史对艾滋病药物治疗有着比较明显的影响，在使用艾滋病治疗方案期间不应该同时进行静脉用药的操作。

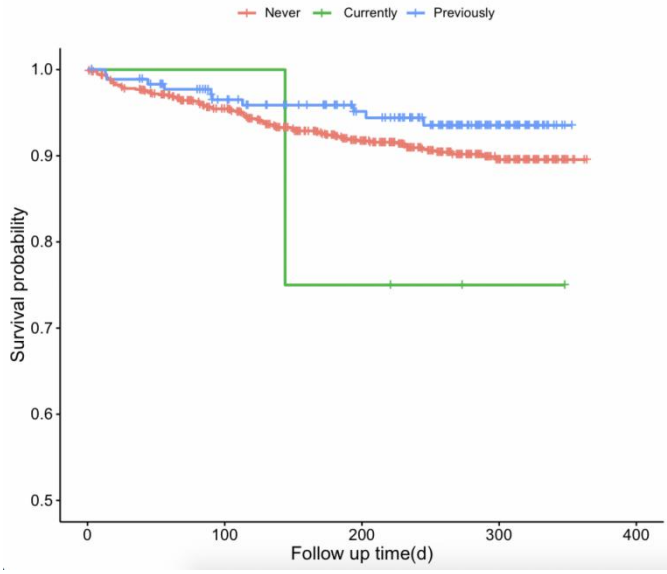

图 9 静脉用药史分组生存曲线

（七）自变量：是否患有血友病

血友病为一组遗传性凝血功能障碍的出血性疾病，其共同的特征是活性凝血活酶生成障碍，凝血时间延长，终身具有轻微创伤后出血倾向。血友病与艾滋病看起来是两种不同类型的疾病，但也有可能具有一定的联系性。为了探究一二，我们绘制了生存曲线，得到结果如图 10。

由结果得到，患有血友病的患者在 100～200 天期间生存率要高于未患血友病的患者，而在 200 天后，血友病患者下降的较多。未患血友病的人生存概率一直下降，波动性较患血友病人群稳定，最终的生存概率要更高一些。两组生存概率都在 85%以上。

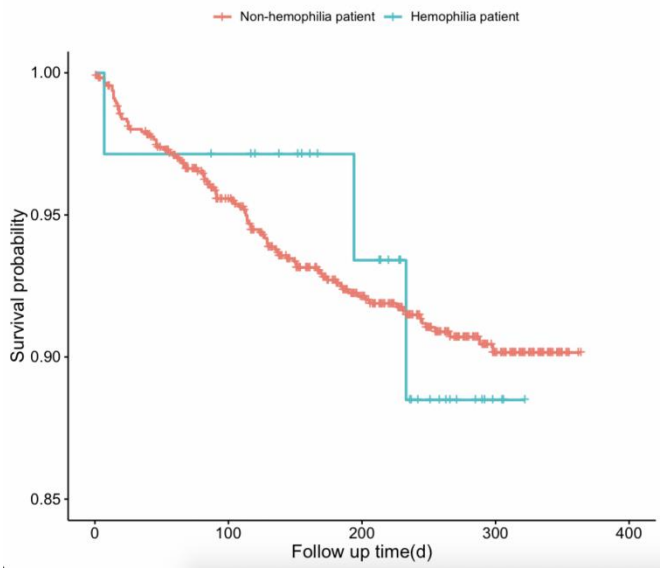

图 10 是否患有血友病分组生存曲线

（八）自变量：基准 CD4 细胞计数、之前几个月使用 ZDV 和年龄

齐多夫定(ZDV)也是一个治疗艾滋病的药物，试验中的患者都至少接受过 3 个月的齐多夫定治疗且每立方毫米的 CD4 细胞数不超过 200 个。

为了更加了解数据的情况，我们对数据每位患者的基准 CD4 细胞计数、之前使用 ZDV 的情况以及年龄做了分布图，将数据可视化，可以更加直观的观察到数据的基本情况。可以发现，大多数患者的 CD4 细胞数都在 0～200cm^3 之间，少数在 200cm^3 以上。接受的齐多夫定治疗都在 0～50 个月之间，极少数在 300 个月，而患者年龄集中在 30～50 岁，服从正态分布。

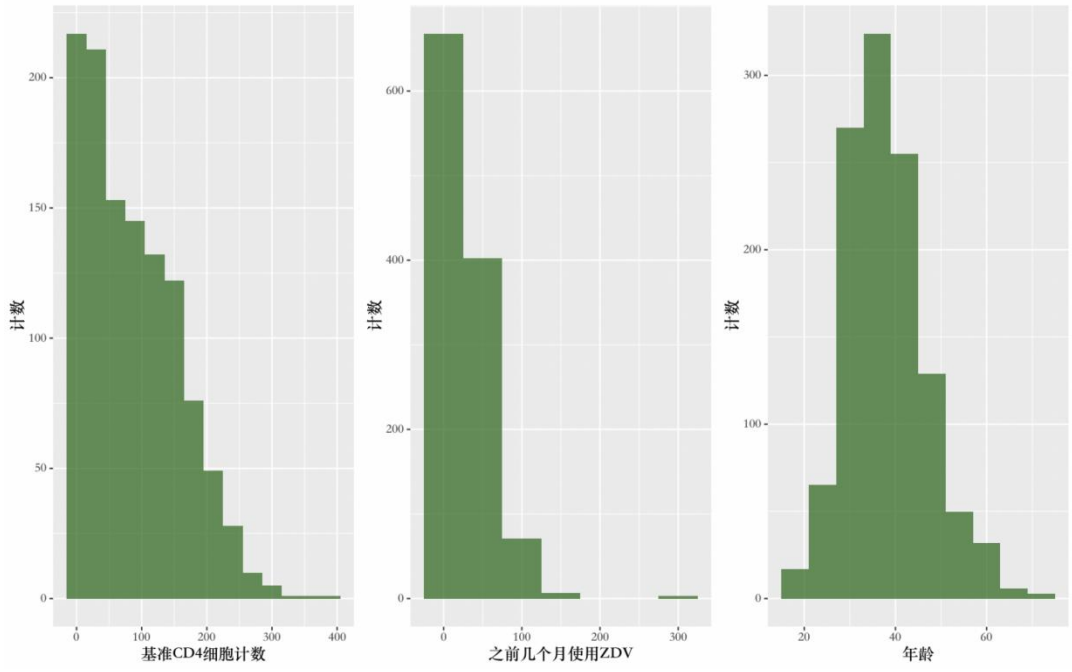

图 11 之基准 CD4 细胞计数、前几个月使用 ZDV 和年龄分布图

四、广义 Gehan-Wilcoxon 检验

在生物学、医学等研究中常常需要对生存函数进行比较。从生存函数曲线的结果中我们发现这些变量取不同值的时候，都可能拥有不同的生存概率，也就是说一个自变量的不同取值可能代表着不同的疾病发展。例如，性别中的男女，女性的生存概率可能要高于男性，但是单一的从生存曲线的图中，我们得不到具有说服力的结论。因此，为了进一步检验是否患者疾病发展规律在变量取值不同时是否存在显著差异，我们进行广义 Gehan-Wilcoxon 检验。

广义 Gehan-Wilcoxon 检验与著名的 Wilcoxon 符号秩检验检验的区别在于它考虑了数据的右删失情况，这也正好符合本文数据的需求，因此，我们利用 Gehan-Wilcoxon 检验来解决所求解问题。

假设备择假设 H_1：存在 $t>0$，使得$S_1(t)>S_2(t)$。我们有两组数据，分别为：$x_1,\ldots,x_{r_1}\quad x_{r_1+1}^{+},\ldots,x_{n_1}^{+}$ 和 $y_1,\ldots,y_{r_2}\quad y_{r_2+1}^{+},\ldots,y_{n_2}^{+}$。定义:

$$U_{ij}=\begin{cases}+1 & \text{if } x_i>y_j \text{ or } x_i^{+}\geqslant y_j\\ 0 & \text{if } x_i=y_j \text{ or } x_i^{+}<y_j \text{ or } y_j^{+}<x_i \text{ or } (x_i^{+},\ y_j^{+})\\ -1 & \text{if } x_i<y_j \text{ or } x_i\leqslant y_j^{+}\end{cases}$$

$$W=\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}U_{ij}$$

我们要计算的统计量则为 $Z=W/\sqrt{\mathrm{Var}(W)}$ ，其中 $\mathrm{Var}(W)=\dfrac{n_1n_2\sum_{i=1}^{n_1+n_2}U_i^2}{(n_1+n_2)(n_1+n_2-1)}$ 利用 R 语言进行编程，得到数据结果。

表 2　Gehan-Wilcoxon 检验结果

自变量	W_G	V	Z
治疗中止指征	14882	23459492	3.072571
筛查时 CD4 层	30786	22139888	6.542829
性别	342	13472290	0.09317629
是否患有血友病	261	2766707	0.156913

由于 $Z_{0.95}$=1.64，因此只有治疗中止指征和筛查时 CD4 层拒绝原假设，意味着包含 IDV 的组比不包含 IDV 的组的疾病发展速度更慢一些；CD4 > 50 比 CD4 <= 50 的组的疾病发展速度更慢一些；性别和是否患血友病之间的治疗效果没有显著性不同。

五、对数秩检验

对数秩检验与 Gehan-Wilcoxon 检验的假设检验一致，它也可以来比较两条生存曲线。由于一些变量的水平不止两个，为了使得编程更简便，我们采用对数秩检验对两水平以上的变量进行检验，得到结果如下：

表 3　对数秩检验结果

自变量		P 值
种族/民族	非西班牙裔白人 Vs 非西班牙裔黑人	0.65
	非西班牙裔白人 Vs 西班牙裔	0.68
	非西班牙裔白人 Vs 亚裔，太平洋岛民	0.30
	非西班牙裔白人 Vs 亚裔，太平洋岛民	0.96
	非西班牙裔黑人 Vs 西班牙裔	0.48
	非西班牙裔黑人 Vs 亚裔，太平洋岛民	0.23
	非西班牙裔黑人 Vs 美洲印第安人	0.96
	西班牙裔 Vs 亚裔，太平洋岛民	0.38
	西班牙裔 Vs 美洲印第安人	0.96
	亚裔，太平洋岛民 Vs 美洲印第安人	0.65
静脉用药史	之前 vs 现在	0.23
	之前 vs 从未	0.23
	现在 vs 从未	0.31
生活质量评分	70 vs 80	0.067
	70 vs 90	5.9e-07
	70 vs 100	2.2e-09
	80vs 90	5.9e-05
	80 vs 100	5.9e-07
	90 vs 100	0.102

可以得到结论：种族和静脉用药史之间没有显著差异，而生活质量评分之间存在明显差异，生活质量 70 的人群与 90 和 100 的人群之间病情发展存在差异；生活质量 80 的人群与 90 和 100 的人群之间病情发展存在差异。我们因此可以判断人们不同的生活质量对艾滋病病情的发展存在很重要的影响。

六、Cox 比例风险模型

（一）单变量 Cox 模型

我们首先计算所有这些变量的单变量 Cox 分析。然后使用重要的变量来拟合多变量 Cox 分析，以描述这些因素如何共同影响生存。

表 4　单变量 Cox 模型结果

变量	回归系数	Wald 检验	p 值
治疗中止指征	-0.68	10	0.0014
筛查时 CD4 层	-1.3	37	1.40E-09
性别	-0.079	0.08	0.78
种族/民族	0.1	0.83	0.36
静脉用药史	-0.22	1.8	0.18
是否患有血友病	0.02	0	0.97
生活质量评分	-0.075	40	2.10E-10
基准 CD4 细胞计数	-0.016	42	9.70E-11
之前几个月使用 ZDV	-0.0025	0.43	0.51
年龄	0.02	3.5	0.061

表 4 显示了每个变量相对于总生存率的回归系数，Wald 检验和 p 值，我们通过单独的单变量 Cox 回归评估每个因素。其中回归系数的符号可以判断每个变量内部之间的联系。正号表示该变量值为不良风险因素，与受试者高事件发生率有关，因此预后较差。例如，性别的回归系数为-0.079，这说明女性比男性具有更低的死亡风险，即女性病情发展的比男性要慢。

从结果得到治疗中止指征、筛查时 CD4 层、生活质量评分、基准 CD4 细胞计数在 5%的显著性水平下具有极高的统计学意义，年龄在 10%的显著性水平下也显著，而其他均不显著。

（二）多元 Cox 模型

为了描述这些因素如何共同影响生存。我们将显著的变量选择进行多元 Cox 回归分析。最终将治疗中止指征、筛查时 CD4 层、静脉用药史、生活质量评分、基准 CD4 细胞计数和年龄 6 个因素纳入多元模型。得到结果如下。

表 5　多元 Cox 模型结果

	风险比	左边界	右边界	P 值	基准组
治疗中止指征：治疗组（包括 IDV）	0.5	0.33	0.76	0.001	对照组
筛查时 CD4 层：：CD4 > 50 组	0.46	0.24	0.87	0.016	CD4 <=50 组
静脉用药史：当前	3.23	0.43	24.45	0.256	没有用药史
静脉用药史：以前	0.51	0.26	0.98	0.045	
生活质量评分：80 分	0.68	0.33	1.41	0.299	
生活质量评分：90 分	0.33	0.16	0.68	0.003	70 分
生活质量评分：100 分	0.21	0.09	0.48	0	
基准 CD4 细胞计数（取对数）	0.81	0.66	1	0.048	
年龄（取对数）	2.19	0.89	5.41	0.09	

由表 5 得到结果：治疗中止指征的治疗组、CD4 > 50 组、以前拥有静脉用药史、生活质量评分为 90 和 100 分、基准 CD4 细胞计数在 5%的显著性水平下显著，表明这些因素与死亡风险降低之间有很强的关系。

风险比可解释为对危险的倍增效应，因此我们可以比较变量与基准组之间的风险大小。我们发现这几个变量的风险比均为正值，基于基准组，我们得到结论：包括 IDV 的组比不包含治疗的疾病发展风险降低 0.5；CD4 > 50 组比 CD4 <=50 组疾病发展风险降低 0.46；生活质量评分为 80、90、100 的患者与生活质量评分为 70 的患者相比，都可以降低疾病发展的风险；而当前有静脉用药史会增加病情发展的风险。

七、结论与思考

通过上述结果，我们基于一项试验对艾滋病药物疗效影响因素进行生存回归分析。结果表明，的确有一些因素可以影响艾滋病患者的病情发展速度，不同的药物方案也对艾滋病病情控制有着不同的疗效与程度。通过结果我们可以提出一些建议给医学研究者以及艾滋病患者，以提高控制艾滋病患者病情发展的可能，可以利用 IDV 的方式对患者进行治疗，但要尽力避免与静脉用药同时搭配。同时，患者也要保持一颗平常心态，内外兼修，从而可以更好的达到药物治疗的目的。

除此之外，本文的分析也引发了我们的一些思考。本文的数据仍然存在缺陷，数据的右删失情况严重，从潜伏到发现症状再到死亡的样本仅仅只有几十例，因此可以考虑延长跟踪时间。此外，本文设定由潜伏到确诊为病情发展情况，可以对确诊到死亡阶段再次进行深入研究，使得研究范围更加广泛，更具有说服力。

附录 1　R 语言程序代码

```
###加载包
library(dplyr)
library(survival)
library(survminer)
library(survival)
library(ggplot2)
library(gridExtra)
library(plyr) ###这个需要调用 plyr 包中的函数 count，plyr 功能极其强大，可用”数据加工厂“一词形容，它会把一个大问题拆分成可以操作的小问题，独立对小问题进行运算。
library(tidyverse)
library(magrittr)
data <- read.csv('/Users/yuqinhan1229/Desktop/actg320.csv')
summary(data)
table <- as.data.frame(summary(data))
write.csv(table,'/Users/yuqinhan1229/Desktop/1.csv')
###任务一
###考察进入临床研究开始至确诊 AIDs 时长
attach(data)
Surv(time,censor==1)
fit1 <- survfit(Surv(time,censor==1) ~ 1,  # 创建生存对象
                data = data) # 数据集来源
fit1 # 查看拟合曲线信息
ggsurvplot(fit1, data = data,
           surv.median.line = "hv",
           break.x.by = 50,
           censor.size=0,
           ylim = c(0.88,1)) # 增加中位生存时间


###考察进入临床研究开始至死亡时长
Surv(time_d,censor_d==1) ##对于患者，创建生存对象，censor_d 为 0 的是右删失，为 1 的死亡
fit2 <- survfit(Surv(time_d,censor_d==1) ~ 1,  # 创建生存对象
                data = data) # 数据集来源
```

```
ggsurvplot(fit2, data = data,
           surv.median.line = "hv",
           ylim = c(0.95,1)) #  增加中位生存时间

###任务 2
####考察不同属性的患者，如性别不同、种族不同、KPS 得分、stata2（进入研究时 CD4 水平高低）、以及治疗方案有无 IDV 时等等，疾病发展速度生存曲线，比较是否存在比较明显的差异。
####性别不同
fit_s <- survfit(Surv(time,censor==1) ~ sex,   #  创建生存对象
                 data = data) #  数据集来源
ggsurvplot(fit_s, #  创建的拟合对象
           data = data,   #  指定变量数据来源
           surv.median.line = "hv",   #  添加中位生存时间线
           xlab = "Follow up time(d)", #  指定 x 轴标签
           legend.title = "", #  设置图例标题，这里设置不显示标题，用空格替代
           legend.labs = c("male", "female"), #  指定图例分组标签
           break.x.by = 100,
           ylim = c(0.85,1))
           #图例、坐标轴字号


####种族不同
fit_r <- survfit(Surv(time,censor==1) ~ raceth,   #  创建生存对象
                 data = data) #  数据集来源
ggsurvplot(fit_r, #  创建的拟合对象
           data = data,   #  指定变量数据来源
           surv.median.line = "hv",   #  添加中位生存时间线
           xlab = "Follow up time(d)", #  指定 x 轴标签
           legend.title = "", #  设置图例标题，这里设置不显示标题，用空格替代
           legend.labs = c("White Non-Hispanic", "Black Non-Hispanic","Hispanic (regardless of race)",
                           "Asian, Pacific Islander"," American Indian, Alaskan Native"),
           ylim = c(0.85,1)) #  指定图例分组标签
```

```
####KPS 得分
fit_k <- survfit(Surv(time,censor==1) ~ karnof,  # 创建生存对象
                 data = data) # 数据集来源
ggsurvplot(fit_k, # 创建的拟合对象
           data = data,  # 指定变量数据来源
           surv.median.line = "hv",  # 添加中位生存时间线
           xlab = "Follow up time(d)", # 指定 x 轴标签
           legend.title = "",
           legend.labs = c("70(Cares for self)","80(Normal activity with effort)",
                           "90(Normal activity possible)","100(Normal)") ) #
设置图例标题，这里设置不显示标题，用空格替代


####strat2（进入研究时 CD4 水平高低）
fit_st <- survfit(Surv(time,censor==1) ~ strat2
                  ,  # 创建生存对象
                  data = data) # 数据集来源
ggsurvplot(fit_st, # 创建的拟合对象
           data = data,  # 指定变量数据来源
           surv.median.line = "hv",  # 添加中位生存时间线
           xlab = "Follow up time(d)", # 指定 x 轴标签
           legend.title = "",
           legend.labs = c("CD4 <= 50","CD4 > 50"),
           ylim = c(0.75,1))


####治疗方案有无 IV 时
fit_tx <- survfit(Surv(time,censor==1) ~ tx,  # 创建生存对象
                  data = data) # 数据集来源
ggsurvplot(fit_tx, # 创建的拟合对象
           data = data,  # 指定变量数据来源
           surv.median.line = "hv",  # 添加中位生存时间线
           xlab = "Follow up time(d)", # 指定 x 轴标签
           legend.title = "",
           legend.labs = c("Control group "," Treatment includes IDV"),
```

```
                ylim = c(0.85,1))


####静脉用药史
fit_iv <- survfit(Surv(time,censor==1) ~ ivdrug,   # 创建生存对象
                          data = data) #  数据集来源
ggsurvplot(fit_iv, #  创建的拟合对象
                data = data,   #  指定变量数据来源
                surv.median.line = "hv",   #  添加中位生存时间线
                xlab = "Follow up time(d)", #  指定 x 轴标签
                legend.title = "",
                legend.labs = c("Never "," Currently","Previously"),
                ylim = c(0.5,1))


####是否患有血友病
fit_h <- survfit(Surv(time,censor==1) ~ hemophil,   # 创建生存对象
                          data = data) #  数据集来源
ggsurvplot(fit_h, #  创建的拟合对象
                data = data,   #  指定变量数据来源
                surv.median.line = "hv",   #  添加中位生存时间线
                xlab = "Follow up time(d)", #  指定 x 轴标签
                legend.title = "",
                legend.labs = c("Non-hemophilia patient","Hemophilia patient"),
                ylim = c(0.85,1))


####基准 CD4 细胞计数分布图
ggplot(data, aes(x=cd4)) + geom_histogram(binwidth=30,fill="darkgreen",alpha=0.7)
+xlab("基准 CD4 细胞计数")+ylab("计数")+
   theme(text=element_text(family="Songti SC",size=12,face = "bold"))


####之前几个月使用 ZDV 分布图
a1=ggplot(data, aes(x=priorzdv)) +
geom_histogram(binwidth=50,fill="darkgreen",alpha=0.7)   +xlab("之前几个月使用
ZDV")+ylab("计数")+
```

```
  theme(text=element_text(family="Songti SC",size=12,face = "bold"))

####年龄分布图
a2=ggplot(data, aes(x=age)) +
geom_histogram(binwidth=6,fill="darkgreen",alpha=0.7)  +xlab("年龄")+ylab("计数")+
  theme(text=element_text(family="Songti SC",size=12,face = "bold"))

grid.arrange(a1,a2,ncol=2)



###任务 3
data_end = data[data$censor==1,]      ##寿终数据
data_end=data_end[order(data_end$time,decreasing=FALSE),] ##对寿终数据进行排序
data_r = data[data$censor==0,]        ##右删失数据
data=data[order(data$time,decreasing=FALSE),] ###对数据进行排序
##gw 检验计算
gw = as.data.frame(c(1:76))
df = as.vector(data_end$time)
time = df[!duplicated(df)]
gw$time = time   ##不同的死亡寿终数据一共 76 个

###计算 rk
r_k=c()
r=0
for (i in 1:76){
  for (j in 1:1151){
    if(data$time[j]>=gw$time[i]){
      r = r+ 1
    }
  }
  r_k[i]=r
  r=0
}
###计算 dk
```

```
table(data_end$time)   ##统计每个元素出现的频次
d_k = as.data.frame(table(data_end$time)) ###转化成数据框格式
d_k =c(d_k$Freq)


##strat2 的 gw 检验
###计算第一组的 rk1 和 dk1
##计算 r_k1
data1 = data[data$strat2==1,]       ##第一组的数据
data1_end = data1[data1$censor==1,] ##寿终数据
data1_end=data1_end[order(data1_end$time,decreasing=FALSE),] ##对第一组寿终
数据进行排序
data_r1 = data1[data1$censor==0,]          ##第一组右删失数据
r_k1=c()
r=0
for (i in 1:76){
   for (j in 1:712){
      if(data1$time[j]>=gw$time[i]){
         r = r+ 1
      }
   }
   r_k1[i]=r
   r=0
}
##计算 d_k1
table(data1_end$time)   ##统计第一组寿终数据每个元素出现的频次
t1 = as.data.frame(table(data1_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
d_k1<- t0 %>%
   full_join(as.tibble(t1), by = "Var1") %>%
   mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
   select(Var1, Freq)
d_k1 =c(d_k1$Freq)
d_k1 = d_k1 -1
```

```
###计算第二组的 rk1 和 dk1
##计算 r_k2
data2 = data[data$strat2==0,]       ##第一组的数据
data2_end = data2[data2$censor==1,] ##第二组的寿终数据
data2_end=data2_end[order(data2_end$time,decreasing=FALSE),] ##对第二组寿终
数据进行排序
data_r2 = data2[data2$censor==0,]          ##第二组右删失数据
r_k2=c()
r=0
for (i in 1:76){
   for (j in 1:439){
      if(data2$time[j]>=gw$time[i]){
         r = r+ 1
      }
   }
   r_k2[i]=r
   r=0
}
##计算 d_k2
t2 = as.data.frame(table(data2_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
d_k2<- t0 %>%
   full_join(as.tibble(t2), by = "Var1") %>%
   mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
   select(Var1, Freq)
d_k2 =c(d_k2$Freq)
d_k2 = d_k2 -1

##生成 G-W 检验计算表
gw$Dk = d_k
gw$rk = r_k
gw$Dk1 = d_k1
gw$rk1 = r_k1
```

```
gw$Dk2 = d_k2
gw$rk2 = r_k2


Wg=0
for (i in 1:76){
   Wg = gw$Dk2[i]*gw$rk1[i] -gw$Dk1[i]*gw$rk2[i] +Wg
}
n1 = 712
n2 = 439
c = n1*n2/((n1+n2)*(n1+n2-1))
V=0
for (i in 1:76){
   V = gw$Dk[i]*gw$rk[i]*(gw$rk[i]-gw$Dk[i]) + V
}
V= c* V
Z= Wg/sqrt(V)



##tx 的 gw 检验
###计算第一组的 rk1 和 dk1
##计算 r_k1
data1 = data[data$tx==1,]       ##第一组的数据
data1_end = data1[data1$censor==1,] ##寿终数据
data1_end=data1_end[order(data1_end$time,decreasing=FALSE),] ##对第一组寿终
数据进行排序
data_r1 = data1[data1$censor==0,]          ##第一组右删失数据
r_k1=c()
r=0
for (i in 1:76){
   for (j in 1:574){
      if(data1$time[j]>=gw$time[i]){
         r = r+ 1
      }
   }
   r_k1[i]=r
```

```
  r=0
}
##计算 d_k1
table(data1_end$time)   ##统计第一组寿终数据每个元素出现的频次
t1 = as.data.frame(table(data1_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
d_k1<- t0 %>%
  full_join(as.tibble(t1), by = "Var1") %>%
  mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
  select(Var1, Freq)
d_k1 =c(d_k1$Freq)
d_k1 = d_k1 -1

###计算第二组的 rk1 和 dk1
##计算 r_k2
data2 = data[data$tx==0,]       ##第一组的数据
data2_end = data2[data2$censor==1,] ##第二组的寿终数据
data2_end=data2_end[order(data2_end$time,decreasing=FALSE),] ##对第二组寿终
数据进行排序
data_r2 = data2[data2$censor==0,]          ##第二组右删失数据
r_k2=c()
r=0
for (i in 1:76){
  for (j in 1:577){
    if(data2$time[j]>=gw$time[i]){
      r = r+ 1
    }
  }
  r_k2[i]=r
  r=0
}
##计算 d_k2
t2 = as.data.frame(table(data2_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
```

```
## 构造数据
## 合并数据，并相加数值
d_k2<- t0 %>%
  full_join(as.tibble(t2), by = "Var1") %>%
  mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
  select(Var1, Freq)
d_k2 =c(d_k2$Freq)
d_k2 = d_k2 -1

##生成 G-W 检验计算表
gw$Dk = d_k
gw$rk = r_k
gw$Dk1 = d_k1
gw$rk1 = r_k1
gw$Dk2 = d_k2
gw$rk2 = r_k2



Wg=0
for (i in 1:76){
  Wg = gw$Dk2[i]*gw$rk1[i] -gw$Dk1[i]*gw$rk2[i] +Wg
}
n1 = 574
n2 = 577
c = n1*n2/((n1+n2)*(n1+n2-1))
V=0
for (i in 1:76){
  V = gw$Dk[i]*gw$rk[i]*(gw$rk[i]-gw$Dk[i]) + V
}
V= c* V
Z= Wg/sqrt(V)



##tx 的 gw 检验
###计算第一组的 rk1 和 dk1
```

```
##计算 r_k1
data1 = data[data$tx==1,]       ##第一组的数据
data1_end = data1[data1$censor==1,] ##寿终数据
data1_end=data1_end[order(data1_end$time,decreasing=FALSE),] ##对第一组寿终数据进行排序
data_r1 = data1[data1$censor==0,]          ##第一组右删失数据
r_k1=c()
r=0
for (i in 1:76){
   for (j in 1:574){
      if(data1$time[j]>=gw$time[i]){
         r = r+ 1
      }
   }
   r_k1[i]=r
   r=0
}
##计算 d_k1
table(data1_end$time)   ##统计第一组寿终数据每个元素出现的频次
t1 = as.data.frame(table(data1_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
d_k1<- t0 %>%
   full_join(as.tibble(t1), by = "Var1") %>%
   mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~ Freq.x + Freq.y ) )%>%
   select(Var1, Freq)
d_k1 =c(d_k1$Freq)
d_k1 = d_k1 -1

###计算第二组的 rk1 和 dk1
##计算 r_k2
data2 = data[data$tx==0,]       ##第一组的数据
data2_end = data2[data2$censor==1,] ##第二组的寿终数据
data2_end=data2_end[order(data2_end$time,decreasing=FALSE),] ##对第二组寿终数据进行排序
```

```
data_r2 = data2[data2$censor==0,]          ##第二组右删失数据
r_k2=c()
r=0
for (i in 1:76){
   for (j in 1:577){
      if(data2$time[j]>=gw$time[i]){
         r = r+ 1
      }
   }
   r_k2[i]=r
   r=0
}
##计算 d_k2
t2 = as.data.frame(table(data2_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
d_k2<- t0 %>%
   full_join(as.tibble(t2), by = "Var1") %>%
   mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
   select(Var1, Freq)
d_k2 =c(d_k2$Freq)
d_k2 = d_k2 -1

##生成 G-W 检验计算表
gw$Dk = d_k
gw$rk = r_k
gw$Dk1 = d_k1
gw$rk1 = r_k1
gw$Dk2 = d_k2
gw$rk2 = r_k2



Wg=0
for (i in 1:76){
   Wg = gw$Dk2[i]*gw$rk1[i] -gw$Dk1[i]*gw$rk2[i] +Wg
```

```
}
n1 = 574
n2 = 577
c = n1*n2/((n1+n2)*(n1+n2-1))
V=0
for (i in 1:76){
   V = gw$Dk[i]*gw$rk[i]*(gw$rk[i]-gw$Dk[i]) + V
}
V= c* V
Z= Wg/sqrt(V)


##sex 的 gw 检验
###计算第一组的 rk1 和 dk1
##计算 r_k1
data1 = data[data$sex==1,]       ##第一组的数据
data1_end = data1[data1$censor==1,] ##寿终数据
data1_end=data1_end[order(data1_end$time,decreasing=FALSE),] ##对第一组寿终
数据进行排序
data_r1 = data1[data1$censor==0,]          ##第一组右删失数据
r_k1=c()
r=0
for (i in 1:76){
   for (j in 1:951){
      if(data1$time[j]>=gw$time[i]){
         r = r+ 1
      }
   }
   r_k1[i]=r
   r=0
}
##计算 d_k1
table(data1_end$time)   ##统计第一组寿终数据每个元素出现的频次
t1 = as.data.frame(table(data1_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
```

```
d_k1<- t0 %>%
   full_join(as.tibble(t1), by = "Var1") %>%
   mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
   select(Var1, Freq)
d_k1 =c(d_k1$Freq)
d_k1 = d_k1 -1

###计算第二组的 rk1 和 dk1
##计算 r_k2
data2 = data[data$sex==2,]       ##第一组的数据
data2_end = data2[data2$censor==1,] ##第二组的寿终数据
data2_end=data2_end[order(data2_end$time,decreasing=FALSE),] ##对第二组寿终
数据进行排序
data_r2 = data2[data2$censor==0,]          ##第二组右删失数据
r_k2=c()
r=0
for (i in 1:76){
   for (j in 1:200){
      if(data2$time[j]>=gw$time[i]){
         r = r+ 1
      }
   }
   r_k2[i]=r
   r=0
}
##计算 d_k2
t2 = as.data.frame(table(data2_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
d_k2<- t0 %>%
   full_join(as.tibble(t2), by = "Var1") %>%
   mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
   select(Var1, Freq)
d_k2 =c(d_k2$Freq)
```

```
d_k2 = d_k2 -1

##生成 G-W 检验计算表
gw$Dk = d_k
gw$rk = r_k
gw$Dk1 = d_k1
gw$rk1 = r_k1
gw$Dk2 = d_k2
gw$rk2 = r_k2


Wg=0
for (i in 1:76){
   Wg = gw$Dk2[i]*gw$rk1[i] -gw$Dk1[i]*gw$rk2[i] +Wg
}
n1 = 951
n2 = 200
c = n1*n2/((n1+n2)*(n1+n2-1))
V=0
for (i in 1:76){
   V = gw$Dk[i]*gw$rk[i]*(gw$rk[i]-gw$Dk[i]) + V
}
V= c* V
Z= Wg/sqrt(V)

##是否血友病的 gw 检验
###计算第一组的 rk1 和 dk1
##计算 r_k1
data1 = data[data$hemophil==1,]       ##第一组的数据
data1_end = data1[data1$censor==1,] ##寿终数据
data1_end=data1_end[order(data1_end$time,decreasing=FALSE),] ##对第一组寿终
数据进行排序
data_r1 = data1[data1$censor==0,]          ##第一组右删失数据
r_k1=c()
r=0
for (i in 1:76){
   for (j in 1:35){
```

```
    if(data1$time[j]>=gw$time[i]){
      r = r+ 1
    }
  }
  r_k1[i]=r
  r=0
}
##计算 d_k1
table(data1_end$time)   ##统计第一组寿终数据每个元素出现的频次
t1 = as.data.frame(table(data1_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
d_k1<- t0 %>%
  full_join(as.tibble(t1), by = "Var1") %>%
  mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
  select(Var1, Freq)
d_k1 =c(d_k1$Freq)
d_k1 = d_k1 -1

###计算第二组的 rk1 和 dk1
##计算 r_k2
data2 = data[data$hemophil==0,]      ##第一组的数据
data2_end = data2[data2$censor==1,] ##第二组的寿终数据
data2_end=data2_end[order(data2_end$time,decreasing=FALSE),] ##对第二组寿终
数据进行排序
data_r2 = data2[data2$censor==0,]          ##第二组右删失数据
r_k2=c()
r=0
for (i in 1:76){
  for (j in 1:1116){
    if(data2$time[j]>=gw$time[i]){
      r = r+ 1
    }
  }
  r_k2[i]=r
```

```
   r=0
}
##计算 d_k2
t2 = as.data.frame(table(data2_end$time)) ###转化成数据框格式
t0 = as.data.frame(table(gw$time))
##  构造数据
##  合并数据，并相加数值
d_k2<- t0 %>%
   full_join(as.tibble(t2), by = "Var1") %>%
   mutate( Freq = case_when( is.na(Freq.x) ~ Freq.y, is.na(Freq.y) ~ Freq.x, TRUE ~
Freq.x + Freq.y ) )%>%
   select(Var1, Freq)
d_k2 =c(d_k2$Freq)
d_k2 = d_k2 -1

##生成 G-W 检验计算表
gw$Dk = d_k
gw$rk = r_k
gw$Dk1 = d_k1
gw$rk1 = r_k1
gw$Dk2 = d_k2
gw$rk2 = r_k2


Wg=0
for (i in 1:76){
   Wg = gw$Dk2[i]*gw$rk1[i] -gw$Dk1[i]*gw$rk2[i] +Wg
}
n1 = 35
n2 = 1116
c = n1*n2/((n1+n2)*(n1+n2-1))
V=0
for (i in 1:76){
   V = gw$Dk[i]*gw$rk[i]*(gw$rk[i]-gw$Dk[i]) + V
}
V= c* V
Z= Wg/sqrt(V)
```

```
####对数秩检验
##种族的对数秩检验
pairwise_survdiff(Surv(time,censor==1) ~ raceth,data,p.adjust.method = "BH")

##ivdrug 的对数秩检验
pairwise_survdiff(Surv(time,censor==1) ~ ivdrug,data,p.adjust.method = "BH")

##karnof 的对数秩检验
pairwise_survdiff(Surv(time,censor==1) ~ karnof,data,p.adjust.method = "BH")

###cox 等比例风险模型
###单变量 coxph 函数
covariates <- c("tx","strat2","sex","raceth","ivdrug","hemophil","karnof",
                "cd4","priorzdv","age")
univ_formulas <- sapply(covariates,
                        function(x) as.formula(paste('Surv(time, censor==1)~',
x)))

univ_models <- lapply( univ_formulas, function(x){coxph(x, data = data)})
# 提取数据，并制作数据表格
univ_results <- lapply(univ_models,
                       function(x){
                         x <- summary(x)
                         p.value<-signif(x$wald["pvalue"], digits=2)
                         wald.test<-signif(x$wald["test"], digits=2)
                         beta<-signif(x$coef[1], digits=2);#coeficient beta
                         HR <-signif(x$coef[2], digits=2);#exp(beta)
                         HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)
                         HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)
                         HR <- paste0(HR, " (",
                                      HR.confint.lower, "-",
HR.confint.upper, ")")
                         res<-c(beta, HR, wald.test, p.value)
                         names(res)<-c("beta", "HR (95% CI for HR)",
"wald.test",
                                       "p.value")
```

```
                                    return(res)
                                    #return(exp(cbind(coef(x),confint(x))))
                                  })
res <- t(as.data.frame(univ_results, check.names = FALSE))
table<-as.data.frame(res)
write.csv(table,'/Users/yuqinhan1229/Desktop/1.csv')

##多元 cox 模型
fit <- coxph(Surv(time, censor==1) ~
as.factor(tx)+as.factor(strat2)+as.factor(ivdrug)+
                          as.factor(karnof)+log(1+cd4)+log(age),data =    data)
summary(fit)
#  构造表格的列
HR <- round(exp(coef(fit)), 2)
CI <- round(exp(confint(fit)), 2)
P <- round(coef(summary(fit))[,5], 3)

#   给 CI 的列命名
colnames(CI) <- c("Lower", "Higher")

# 将列合并为数据框
table2 <- as.data.frame(cbind(HR, CI, P))
write.csv(table2,'/Users/yuqinhan1229/Desktop/1.csv')

ggsurvplot(survfit(fit), color = "#2E9FDF",
                ggtheme = theme_minimal(),data)
```