



百万医疗险医疗费用通胀 归因分析与赔付预测

——基于机器学习的研究

目录

CONTENTS

- 01 医疗费用通胀现状
- 02 医疗通胀归因分析
- 03 基于机器学习的百万医疗险赔付预测
- 04 结论与建议
- 05 创新点和应用价值

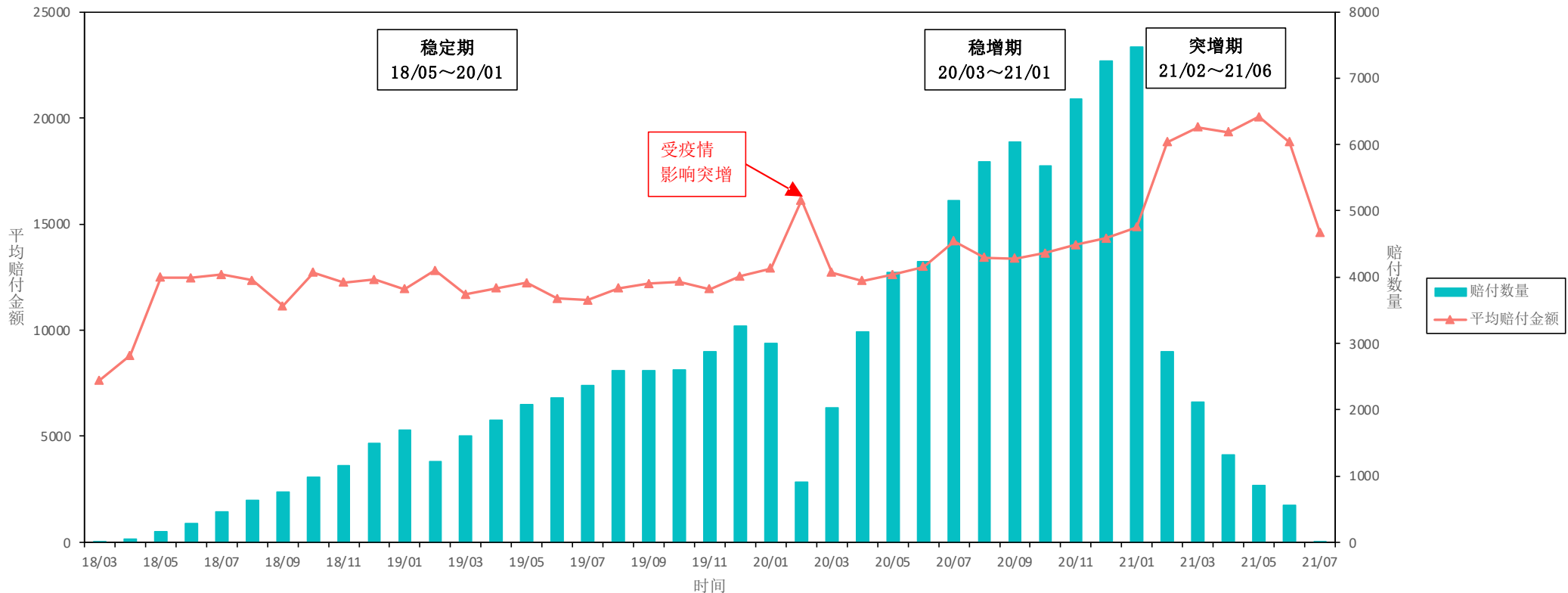
PART 01 ▶

医疗费用通胀现状



1 医疗费用通胀现状

- **平均赔付金额：**从2018年5月的12500元左右上升到2021年6月的19000元左右，**上涨额度约为6500元，上涨幅度超过50%，通胀效应明显。**
- **赔案数量：**自产品开售的2018年3月后一直呈上升趋势直到2021年1月，之后逐渐减少。



平均赔付金额和赔付数量

*以结案时间（理赔时间）划分月份，并去除赔付数量过少的首尾月份。

PART 02 ▶

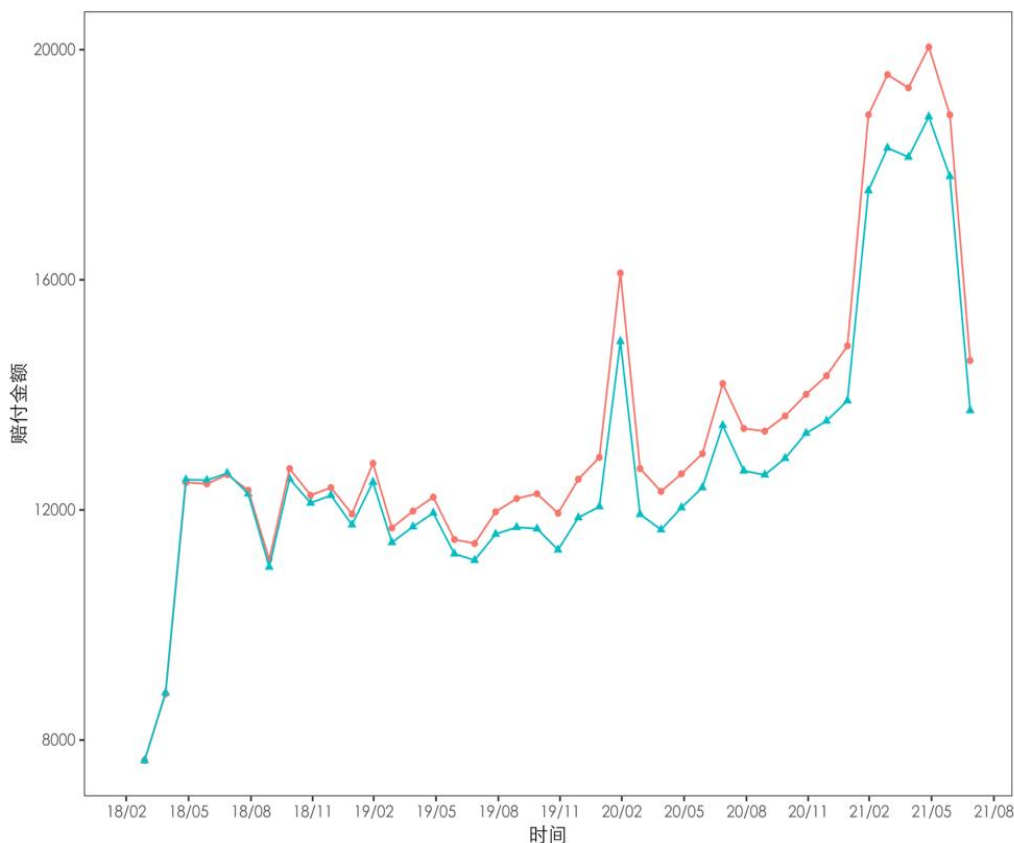
医疗通胀归因分析



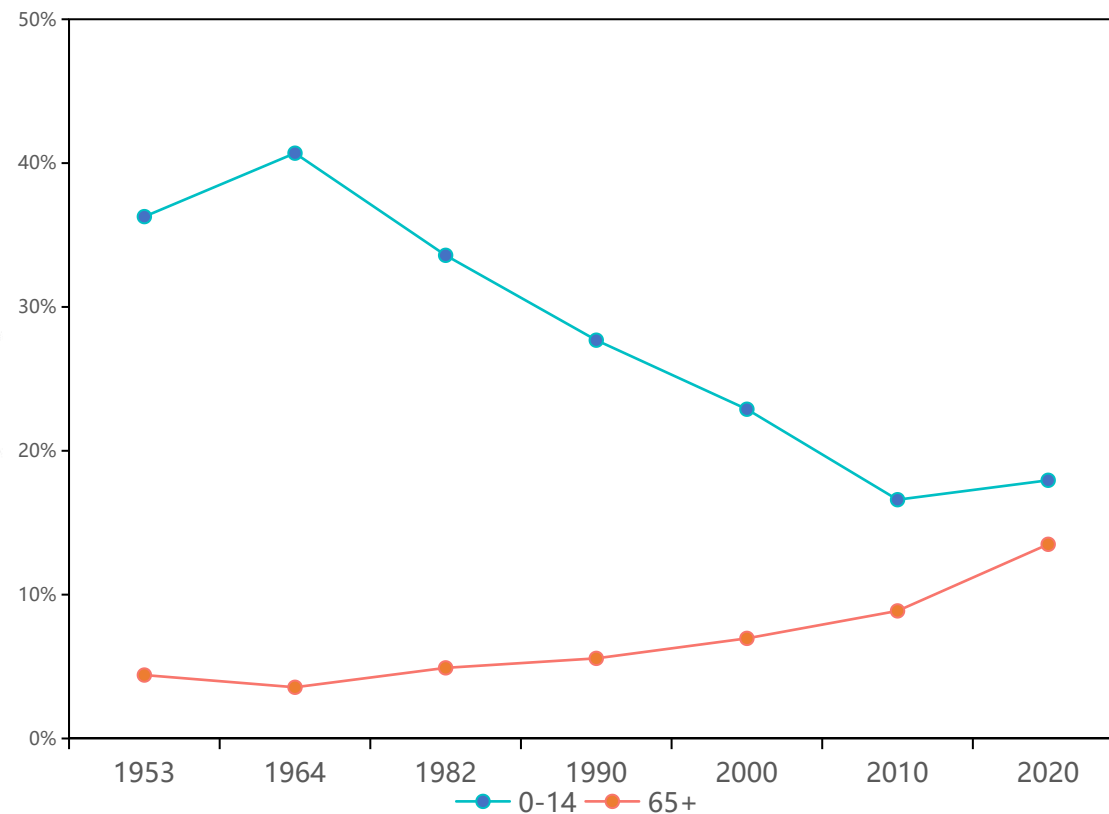
2.1 经济社会变动推高总体医疗通胀

思想自由 兼容并包

- **通货膨胀**：2021年7月的CPI较2018年3月上升了6.3%，年均增长1.9%，物价水平总体在较低通胀水平下波动，直接影响医疗通胀。
- **老龄化**：2020年我国65岁以上人口占比达到13.5%，比上一次人口普查数据8.87%上升了4.63%。预期寿命的增长给医疗保险带来了前所未有的医疗支出压力。



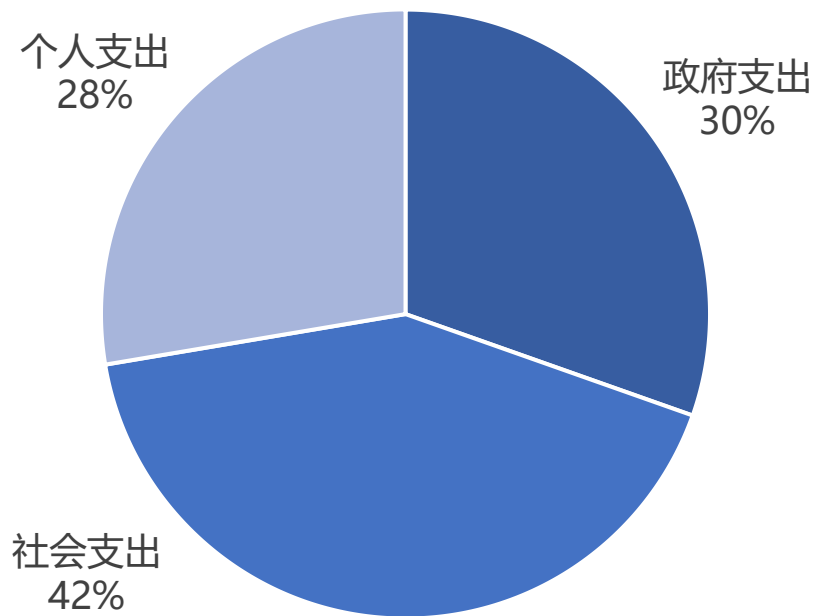
经过CPI平滑后的平均赔付金额



历次人口普查0-14岁和65岁以上人口比重

2.1 经济社会变动推高总体医疗通胀

- **医疗进步：**社会医疗支出占比从2010年的36%升至2020年的42%，说明保险公司在全社会的医疗卫生事业中发挥了越来越大的社会责任。
- **治疗费用上涨：**一方面由于更为先进和昂贵的医疗手段和药物的出现，一方面由于慢性病和恶性肿瘤等疾病的发生率上升和年轻化。



2020年总卫生支出比重

2015~2020年次均住院和门诊医药费

	2015	2020
住院医药费	8,268	10,619
门诊医药费	234	324

2.2 微观风险指标左右实际赔付支出

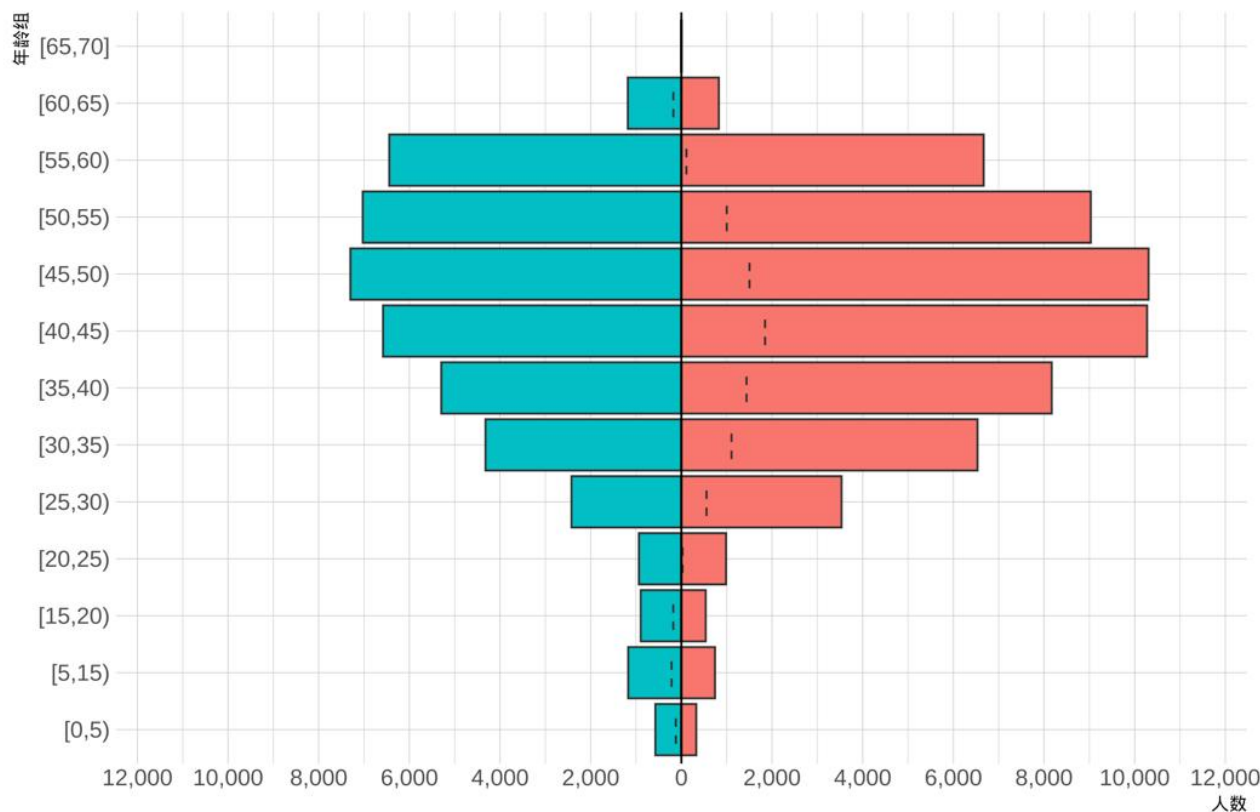
- 我们基于提供的理赔数据，对费用通胀做进一步的分析和归因，观察客户群体和理赔特点等微观因素变化对费用通胀的影响。
- 原始数据包含102095行 x 29列，我们将其归为三类，提取构建出11个指标。

指标类型	指标名称	指标处理
出险人群	参保年龄*	根据保单生效日期-出生日期构建
	性别	男、女
	地区	将27个省份归为4大地区 (西部、中部、东部、东北地区)
	社保类型	有社保、无社保
损失事件	疾病类型	根据ICD-10国际疾病分类标准，将第一位编码为A-R的疾病归类为普通疾病，其余为重大疾病
	赔付类型	疾病、意外、重大疾病医疗、一般医疗
	疾病数量*	根据赔案中疾病种类总数构建
	医院等级	将18种细分等级归为4类 (一级、二级、三级、其他)
	给付类型	正常给付、通融给付
	赔付次数*	根据同一保单对应的赔案数量构建
	原始总费用*	-
业务情况	案件处理速度、赔付比例、通融赔付比例	-

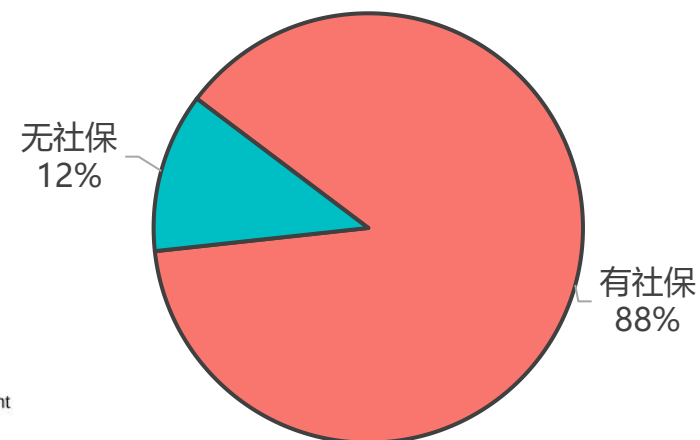
*标记为数值型变量

2.2.1 出险人群结构特征

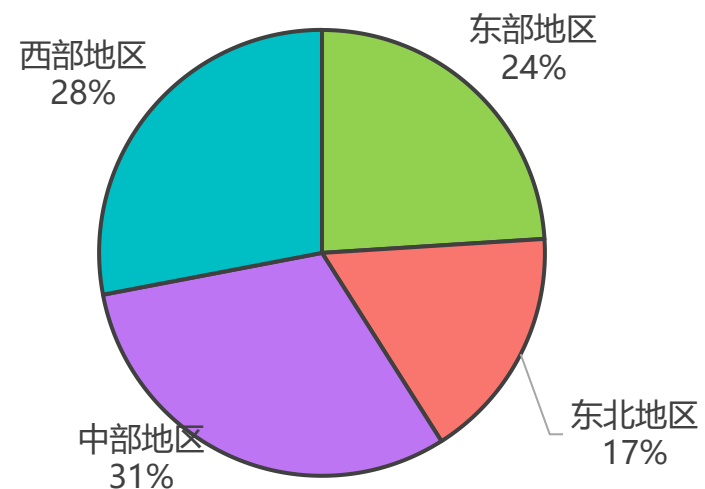
- **性别结构：**女性多于男性，大部分在30岁以上，其中40-55年龄段最为集中，与社会年龄结构有差异。
- **社保类型：**大部分参保人员都同时参加了基本医疗保险，有助于缓解赔付压力。
- **地区分布：**地域较为均匀，有助于分摊赔付风险。



参保人群金字塔



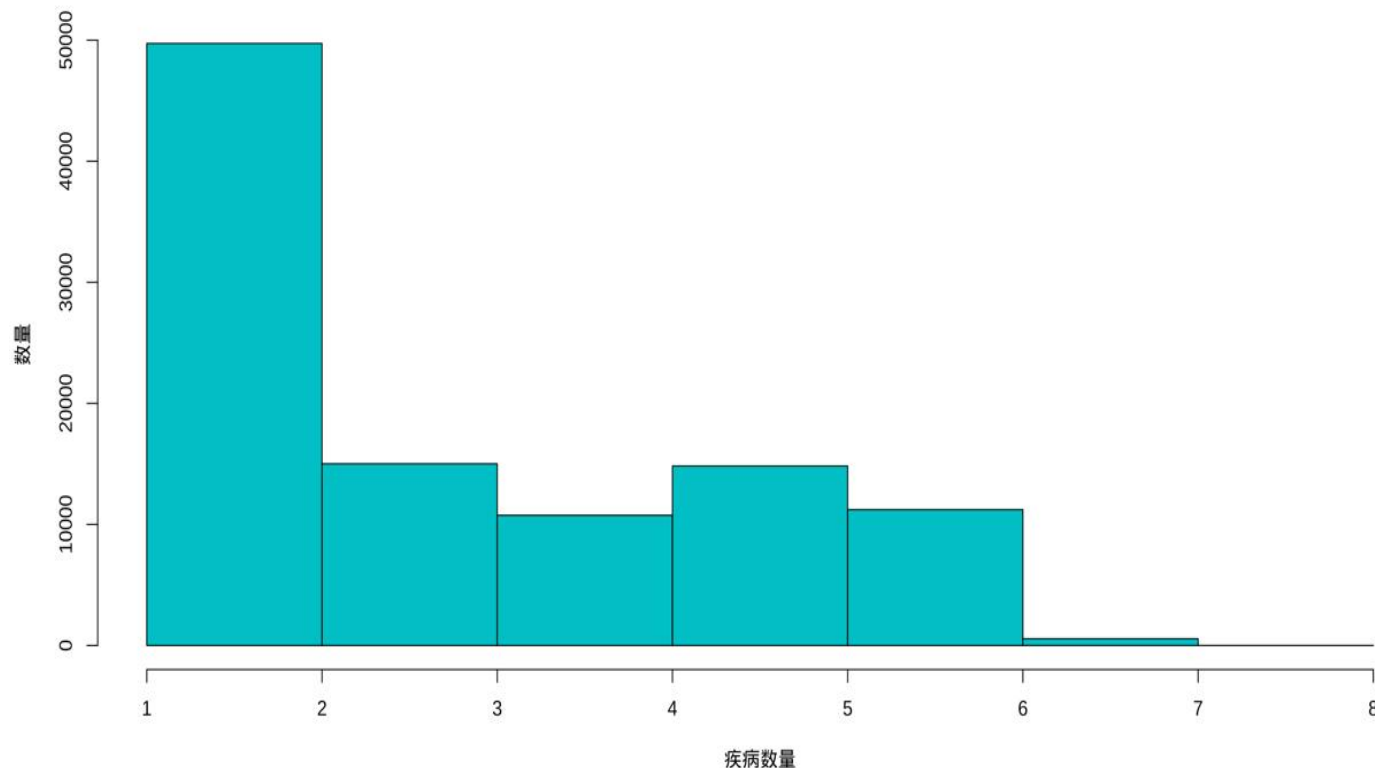
参保人群
社保分布



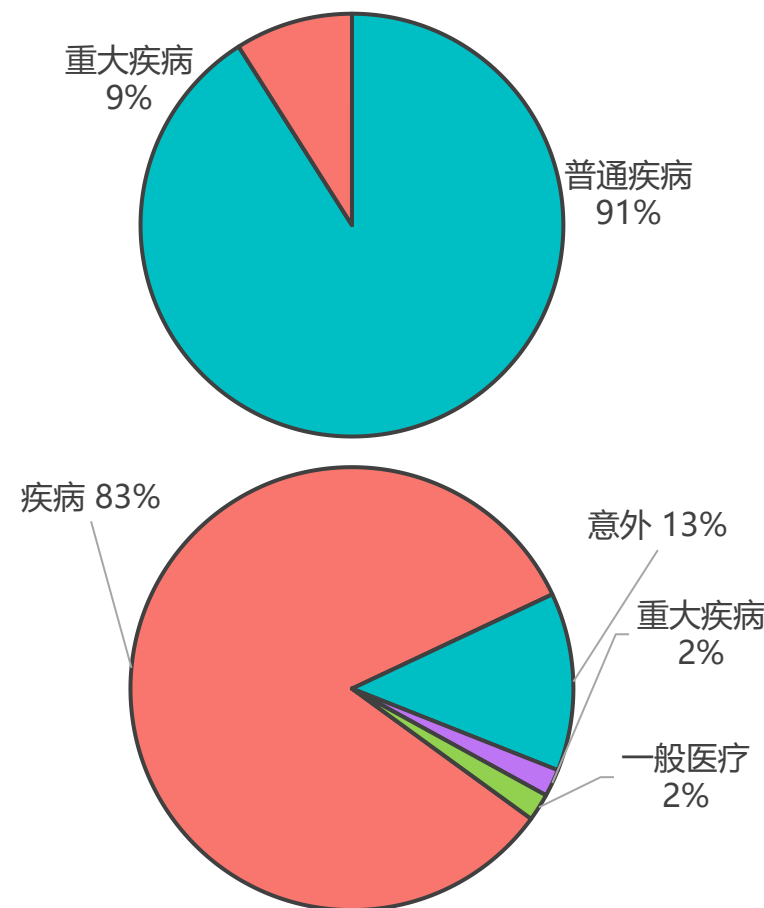
参保人群
地区分布

2.2.2 损失事件结构特征

- **疾病数量：**每次赔付可能对应了1-8个/种疾病，疾病数量越多，可能的赔付越高。
- **疾病类型：**本产品中一般疾病的保险保障设有10000元的免赔额，而重大疾病的保险保障不设免赔额，同时重大疾病有更大概率产生更高额的医疗费用，可能导致更高的赔付支出。
- **赔付类型：**本产品中疾病和意外适用于同样的保障，但意外造成疾病的治疗更具有偶然性，增加赔付的波动。



疾病数量分布特征

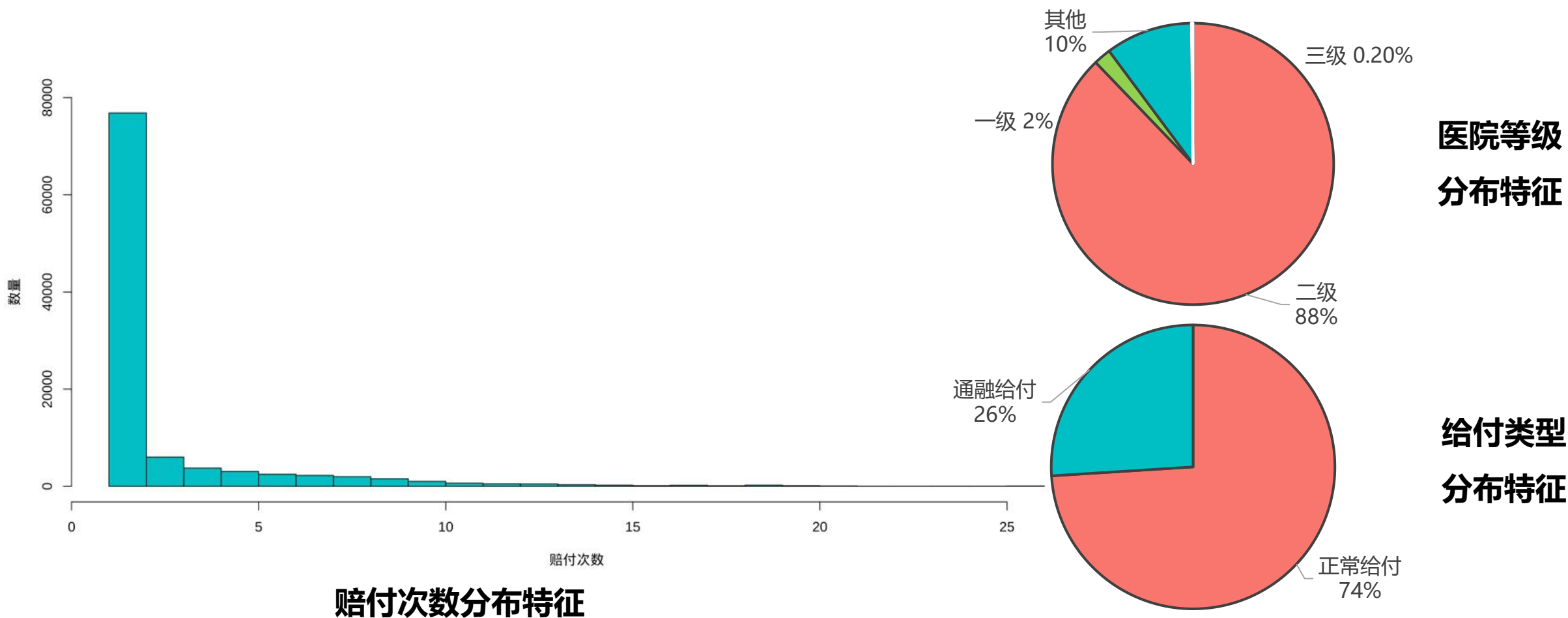


疾病类型
分布特征

赔付类型
分布特征

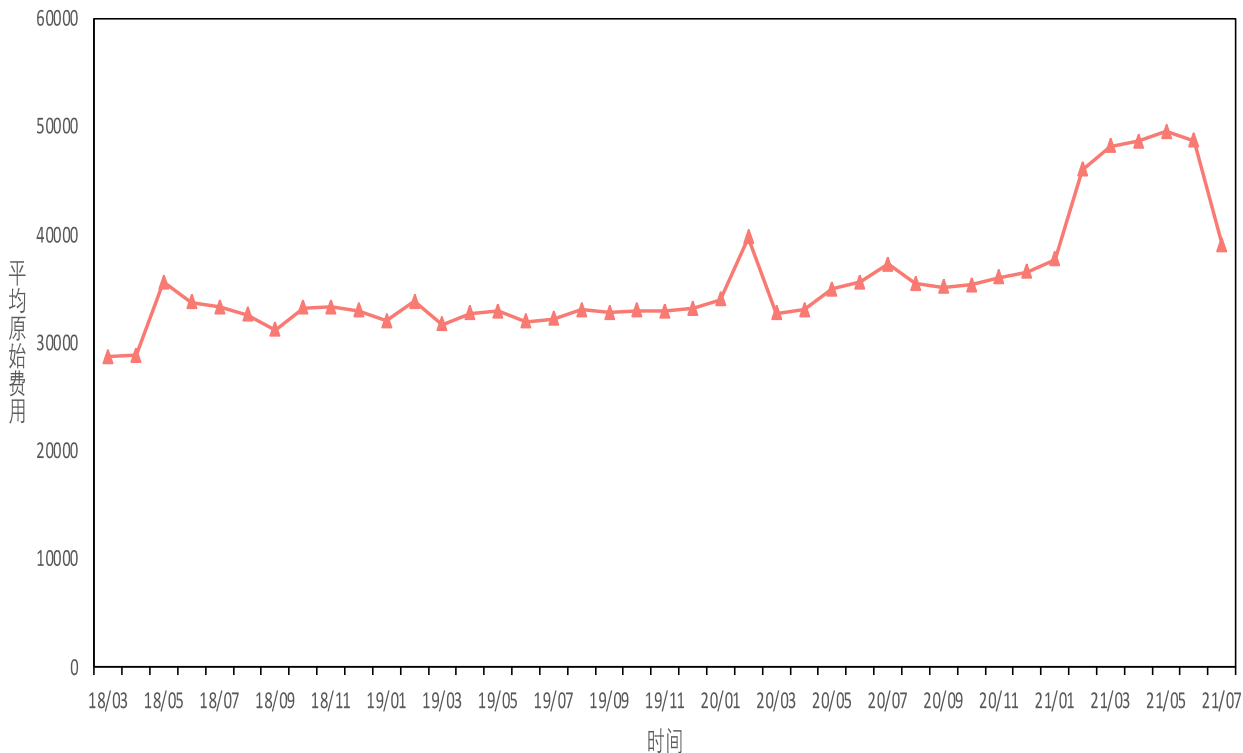
2.2.2 损失事件结构特征

- **赔付次数：**少部分保单在有效期内发生多次赔付，反复出险的保单往往需要特别关注。
- **医院等级：**大部分的赔付发生在二级医院，占比高达88%，通常高等级的医院医疗费用更高。
- **给付类型：**26%的给付属于通融给付，这部分平均赔付额通常小于普通给付，但会增加总支出。

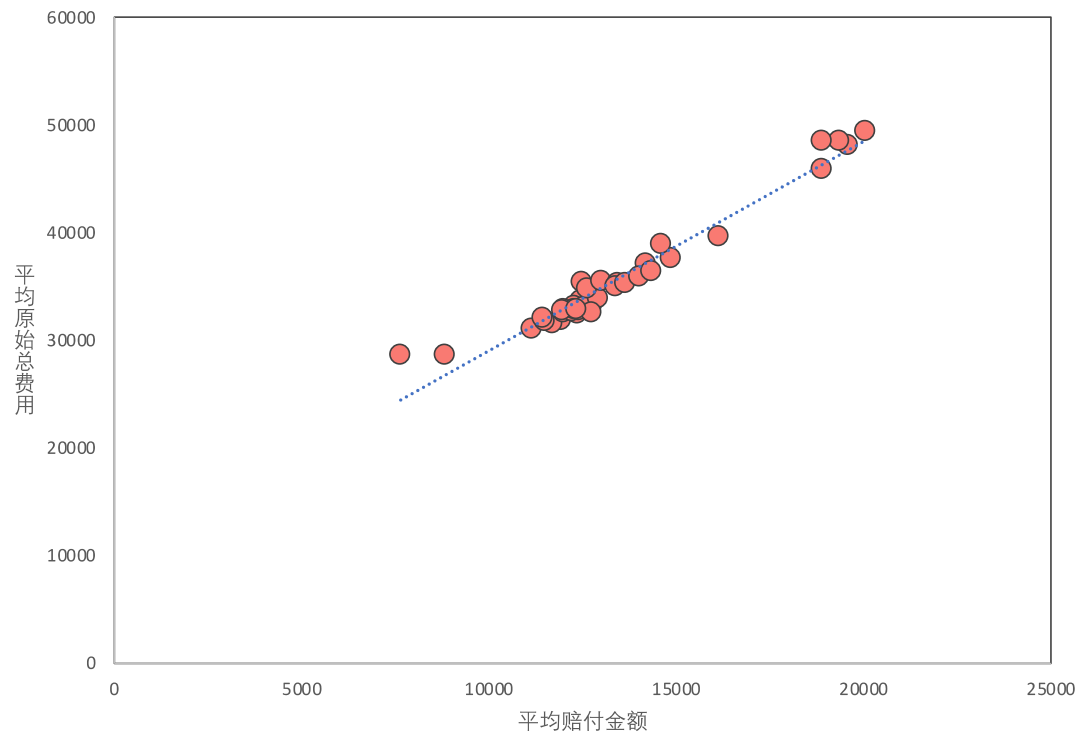


2.2.2 损失事件结构特征

- **原始费用：**平均的医疗总费用在2020年2月前较为稳定，2020年2月开始出现较大波动，之后在2021年2月开始急剧上升，与平均赔付金额趋势较为接近。
- 由于本产品并不设置比例共保，因此原始医疗费用的变化会直接影响到保险公司最终的赔付支出。



平均总费用变化



平均赔付金额与平均总费用的关系

- 为定量分析风险指标的影响，建立风险指标模型，首先将原始数据进行清洗：
 - 费用缺失值补0
 - 剔除六类异常数据（根据费用加总逻辑、日期顺序逻辑、赔案数量逻辑判断）
 - 剔除2021年7月和2018年3月结案赔付的保单（仅11+8=19个，数量过少，不具有代表性）
 - 清洗后剩余98452条数据
- 我们用简单的**Robust回归**建立实际赔付金额与风险指标之间的联系。首先设定回归方程如下：

Compensation

$$\begin{aligned} &= \alpha_1 Age + \alpha_2 Female + \alpha_3 Security + \beta_1 Critical + \beta_2 Class1 \\ &+ \beta_3 Class2 + \beta_4 Class3 + \beta_5 Incident + \beta_6 Type1 + \beta_7 Type2 \\ &+ \beta_8 Payment + \beta_9 Cost + \beta_{10} Times + \alpha_4 East + \alpha_5 Mid + \alpha_6 West \\ &+ \varepsilon \end{aligned}$$

2.3.1 各风险指标平均影响探究

- 回归结果显示，除疾病数量这一变量外，其他变量都与实际赔付金额有显著的相关关系。大部分变量的影响符合我们的预测。
- 值得注意的是，女性对赔付金额有着正向相关关系，意外造成的治疗费用对赔付有负向相关关系，东部地区与赔付金额的负相关关系反而比中西部地区更大。

Robust 回归结果

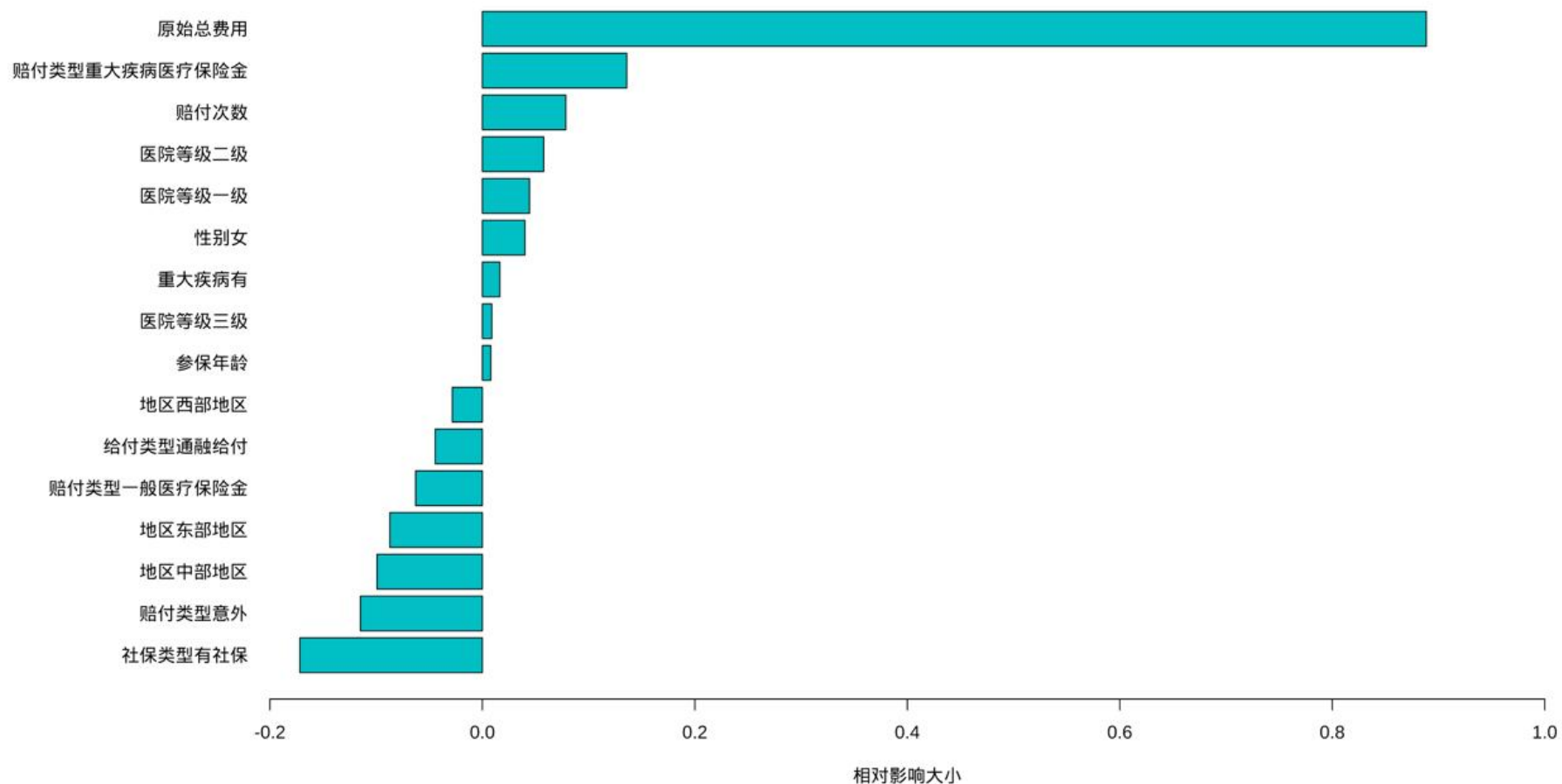
	参数值	标准误	P值
年龄	14.77	2.884	0.000***
性别-女性	860.1	66.10	0.000***
社保类型-有社保	-3686	102.5	0.000***
疾病种类-重大疾病	353.0	150.4	0.019**
医院等级-一级	951.3	245.3	0.000***
医院等级-二级	1238	113.0	0.000***
医院等级-三级	190.4	671.7	0.777
赔付类型-意外	-2463	136.9	0.000***
赔付类型-一般医疗保险金	-1347	210.3	0.000***
赔付类型-重疾医疗保险金	2915	2.399	0.000***
给付类型-通融给付	-953.0	75.58	0.000***
原始总费用	0.4928	0.001	0.000***
赔付次数	551.5	10.93	0.000***
地区-东部地区	-1869	84.98	0.000***
地区-中部地区	-2128	89.28	0.000***
地区-西部地区	-607.9	96.87	0.000***

***P<0.01,**P<0.05,*P<0.1, adjusted R-squared: 0.7773

2.3.1 各风险指标平均影响探究

- 标准化后的各变量影响权重大小如图所示。
- 由于医疗通胀是赔付在时间上的变化，为了进一步查清医疗通胀的微观来源，我们还需要进一步分析这些变量的变化情况。

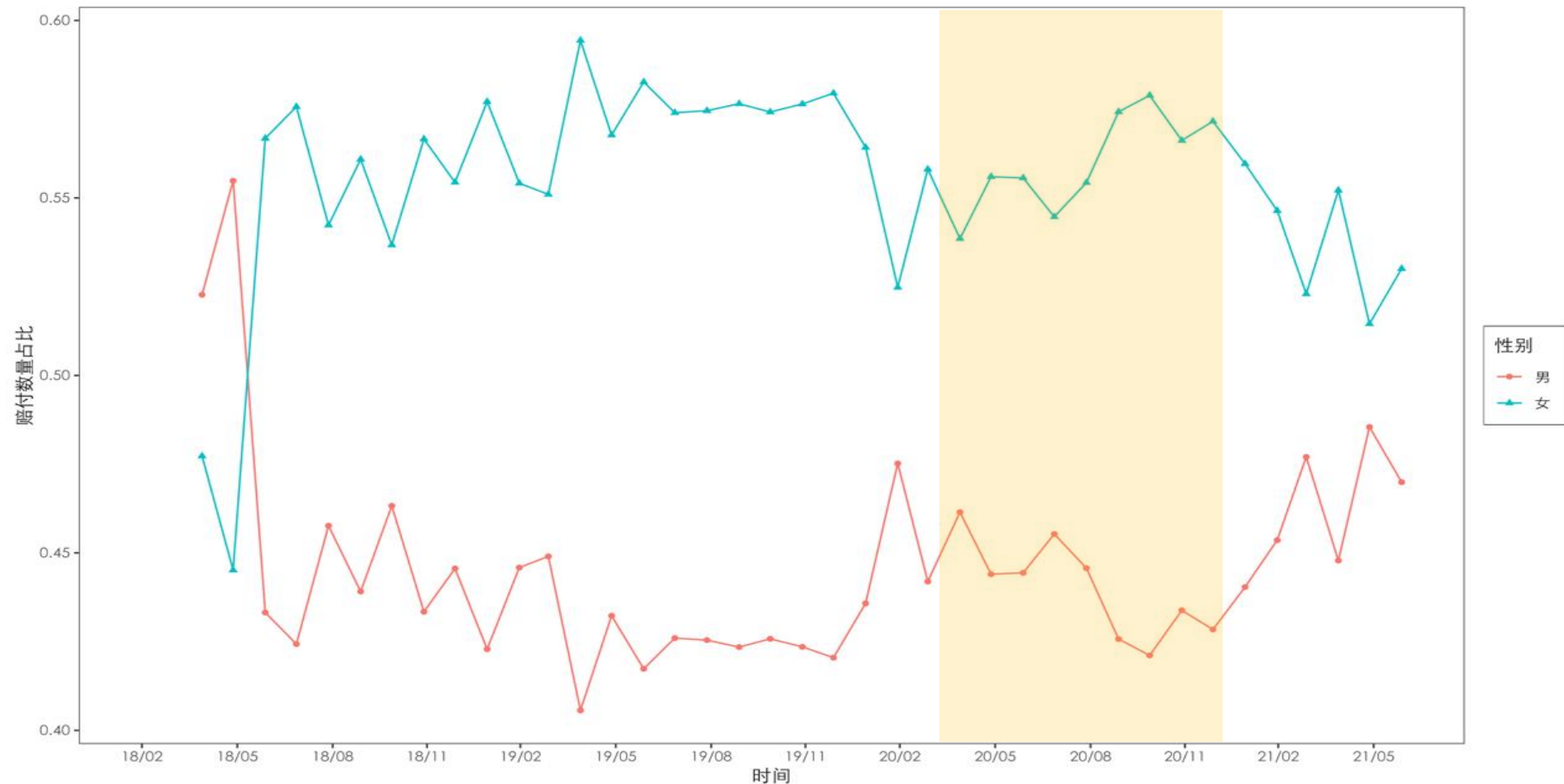
变量影响权重结果



2.3.2 各风险指标时间变动趋势探究

- 自2018年到2020年，在通胀平稳期女性赔案比例稳步上升，而在突增期反而有所下降，女性案件比例上升是平稳期赔付金额通胀的原因之一。

赔案中性别比例变化

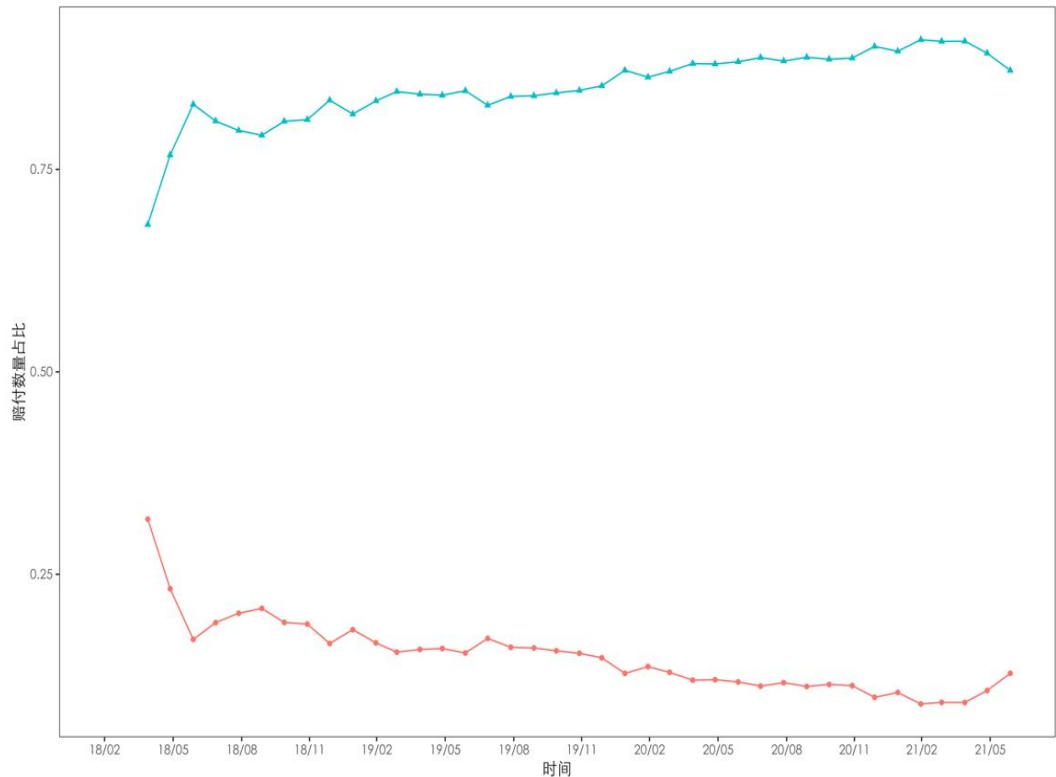


2.3.2 各风险指标时间变动趋势探究

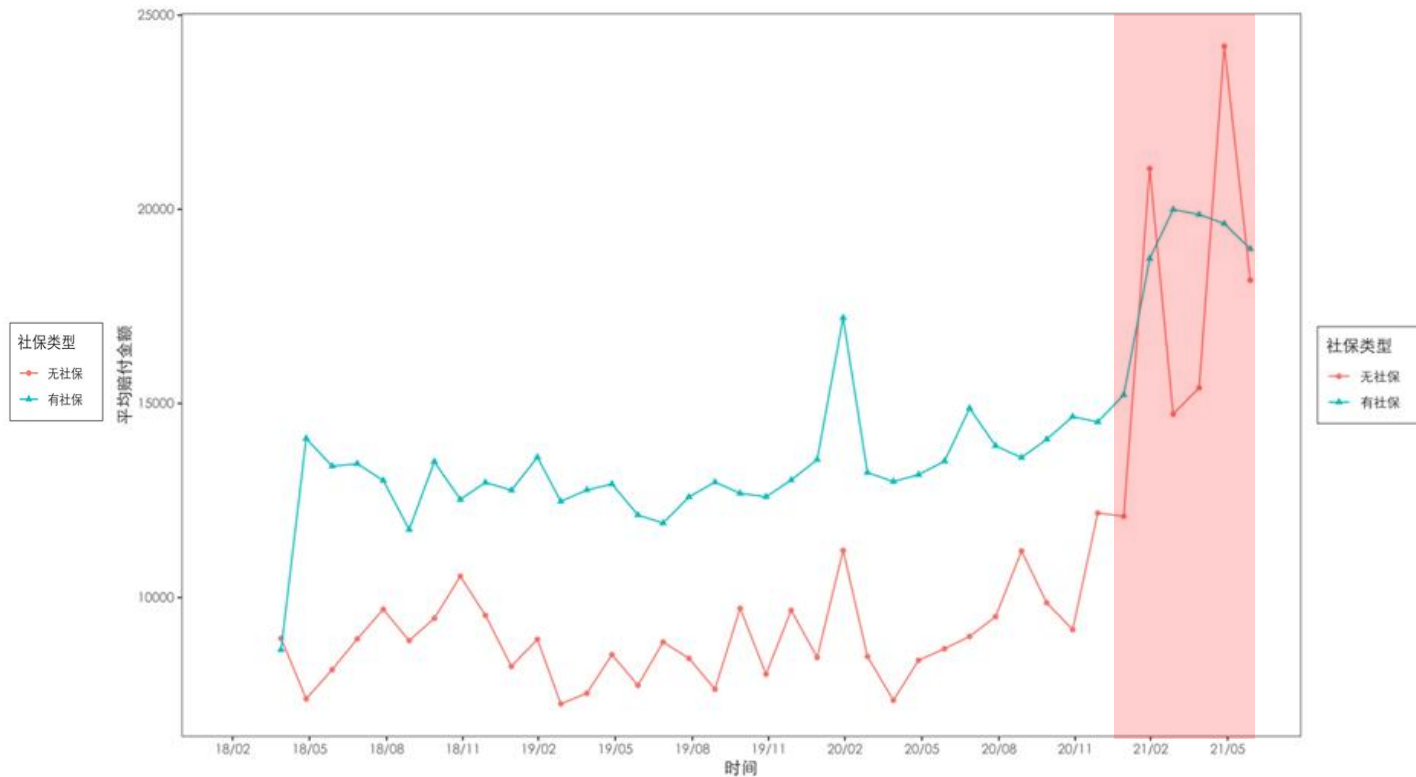
思想自由 兼容并包

- 在社保人群比例基本不变的情况下，通胀突增期没有参加社保人群的赔付金额突然大幅上升，造成了这段时间内的赔付金额通胀。

不同社保类型比例变化



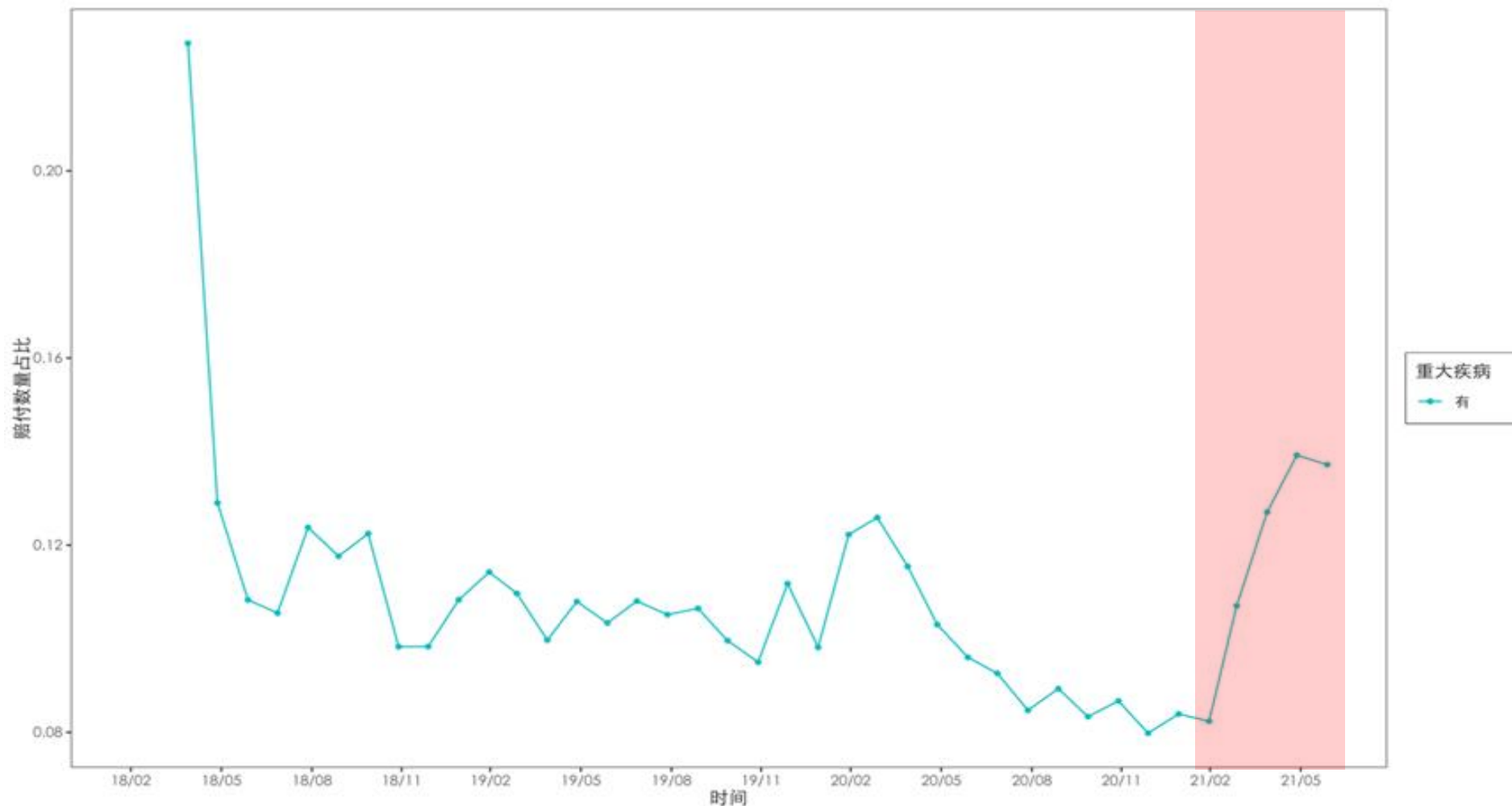
不同社保类型平均赔付金额变化



2.3.2 各风险指标时间变动趋势探究

- 在通胀突增期，重大疾病的赔案比例明显上升至14%左右，推高了平均赔付金额。

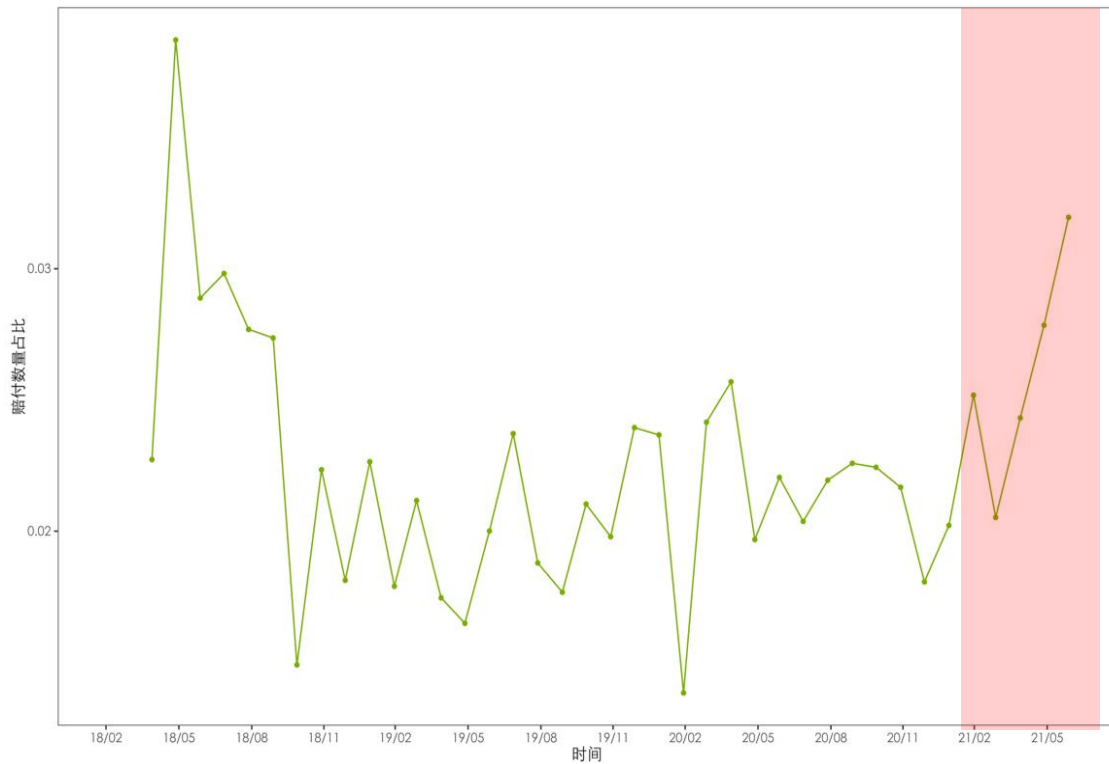
重大疾病比例变化



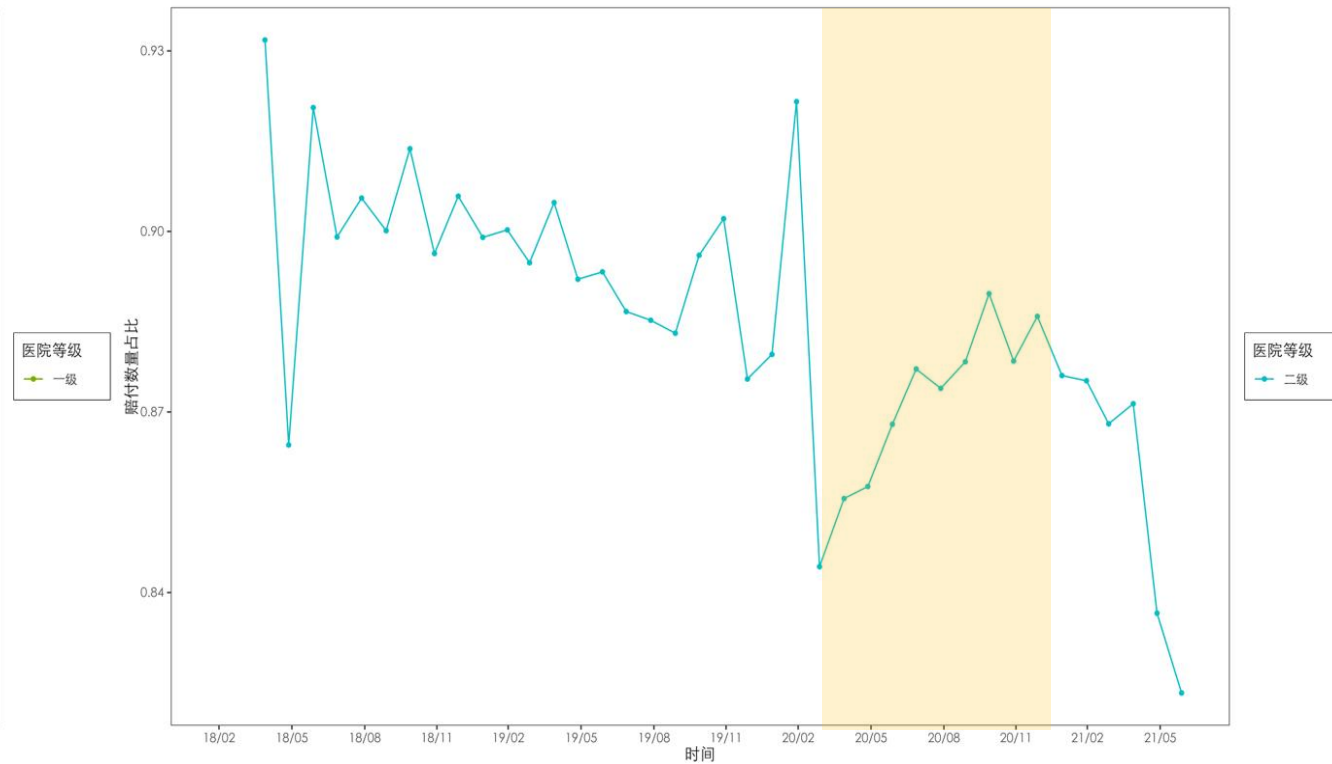
2.3.2 各风险指标时间变动趋势探究

- 二级医院的赔案比例在3年间波动下滑，但在平稳期有所上升，同时一级医院的赔案占比在突增期上升，部分造成了后期的费用上涨。

一级医院比例变化



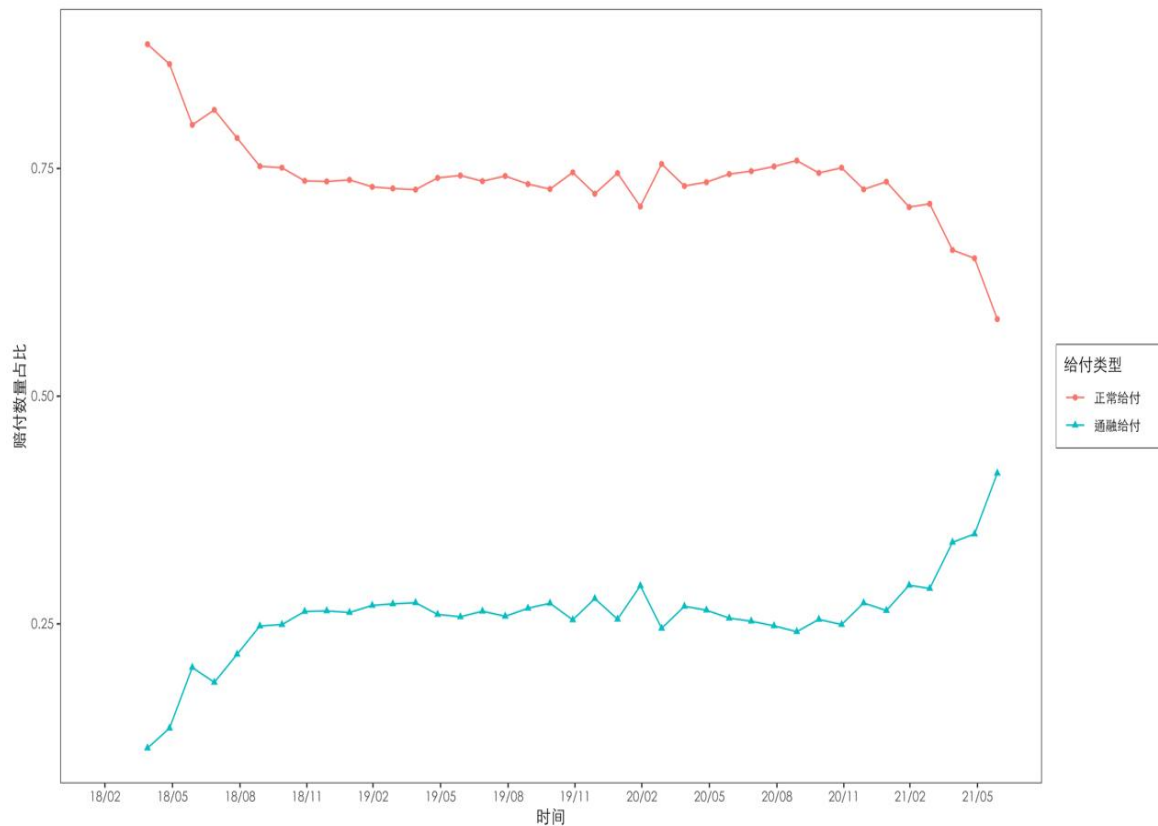
二级医院比例变化



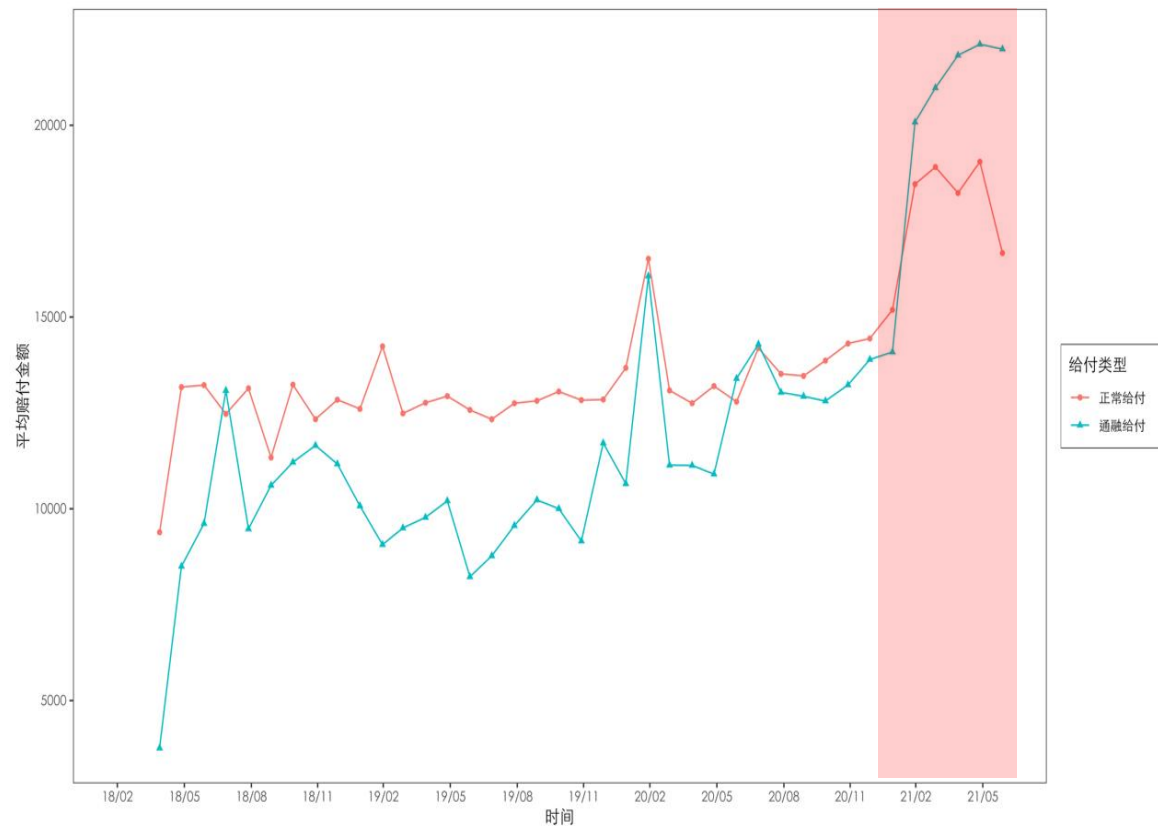
2.3.2 各风险指标时间变动趋势探究

- 给付类型比例在3年内变化较为稳定，而通融给付在后期的平均赔付金额开始上升，是突增期的费用通胀的原因之一。

分给付类型比例变化



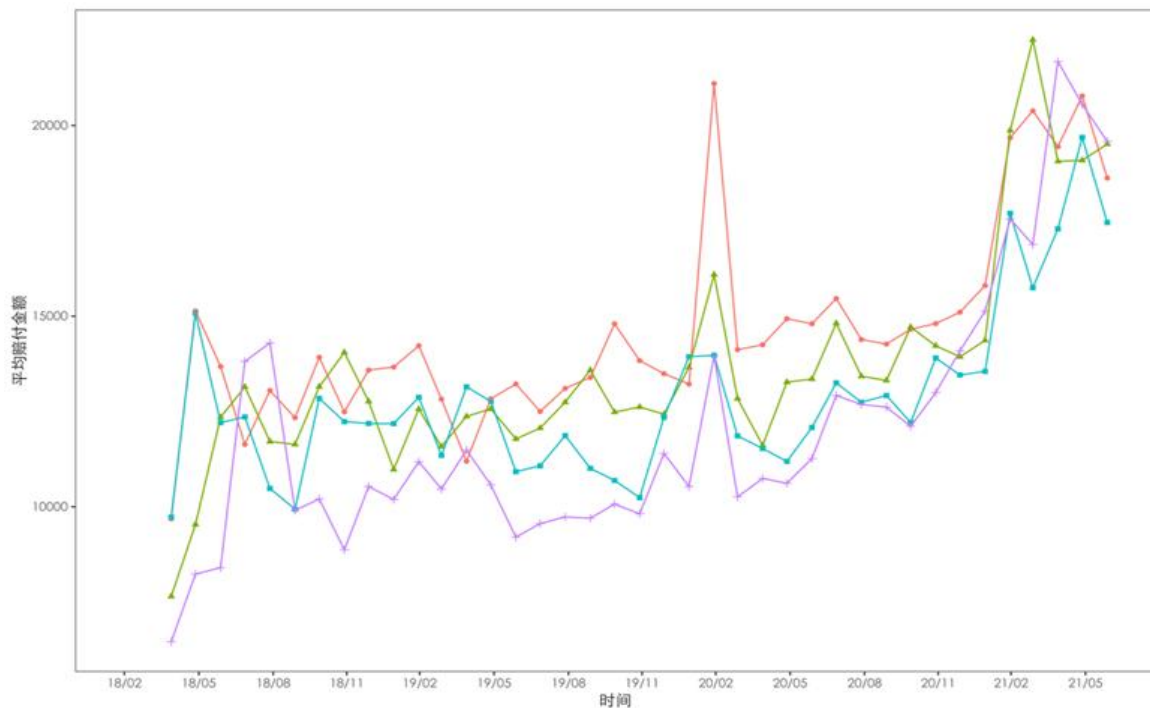
给付类型平均赔付金额变化



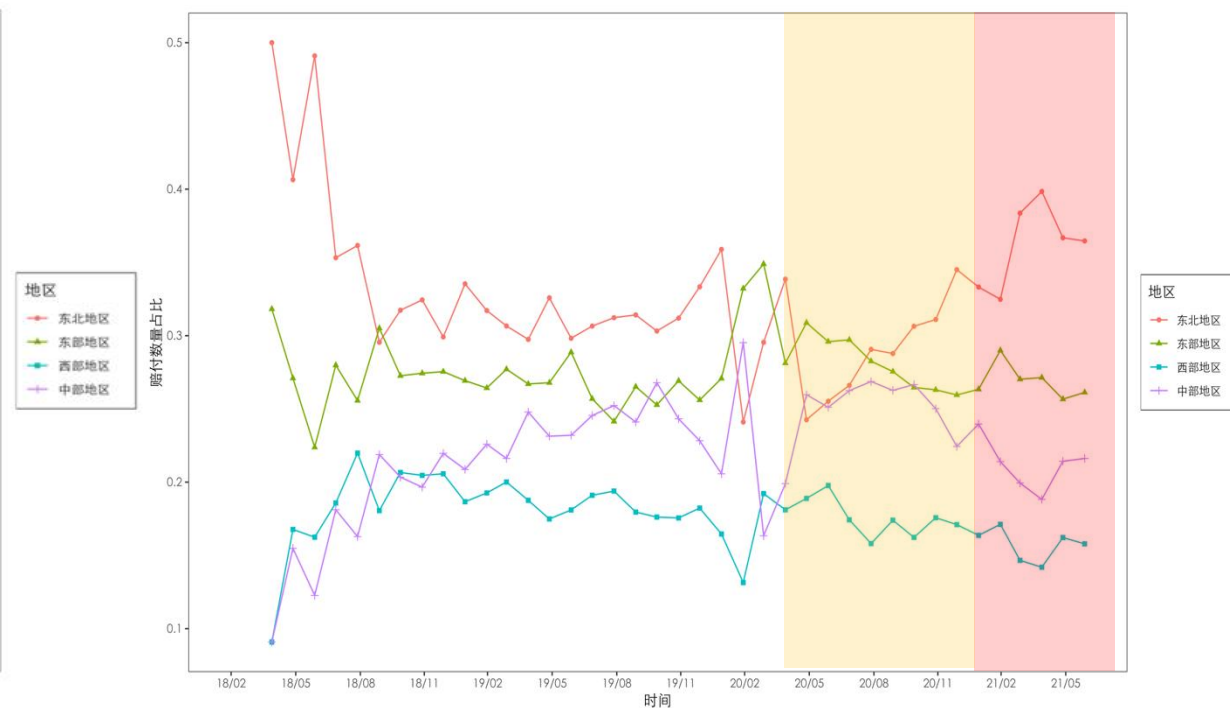
2.3.2 各风险指标时间变动趋势探究

- 分地区来看，东北地区的平均赔付金额更高，这可能是由于东北地区的老龄化程度更高。
- 在3年期内，除去初期赔付数量较少的时间，前期各地区赔付比例相对稳定，而通胀期东北地区的赔付比例突然增大，带动了费用通胀。

分地区平均赔付金额变化



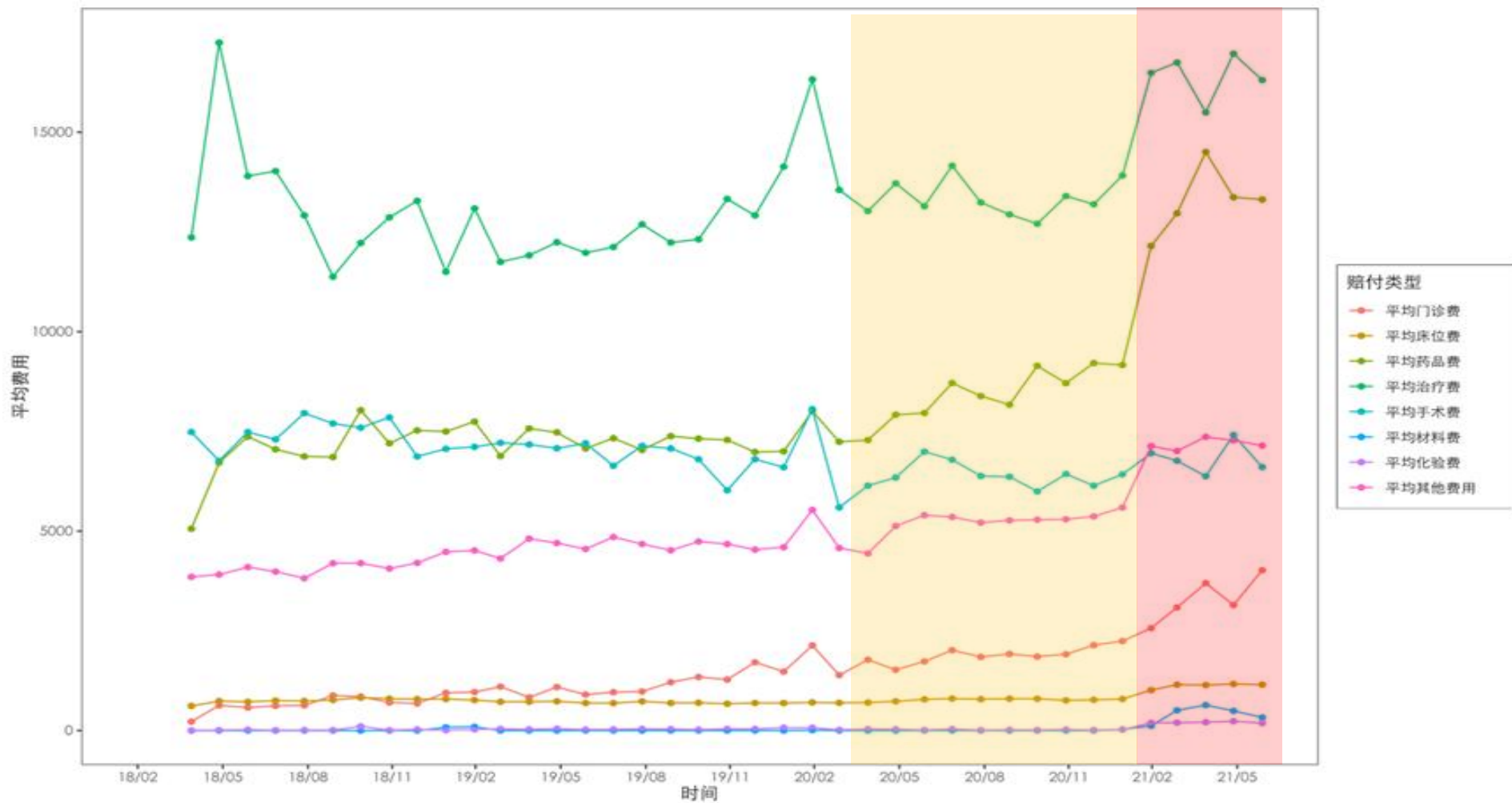
分地区赔付比例变化



2.3.2 各风险指标时间变动趋势探究

- 从医疗费用细分类别来看，通胀期治疗费、药品费的快速上涨带动了整体医疗费用的上涨，进一步推动理赔上涨。

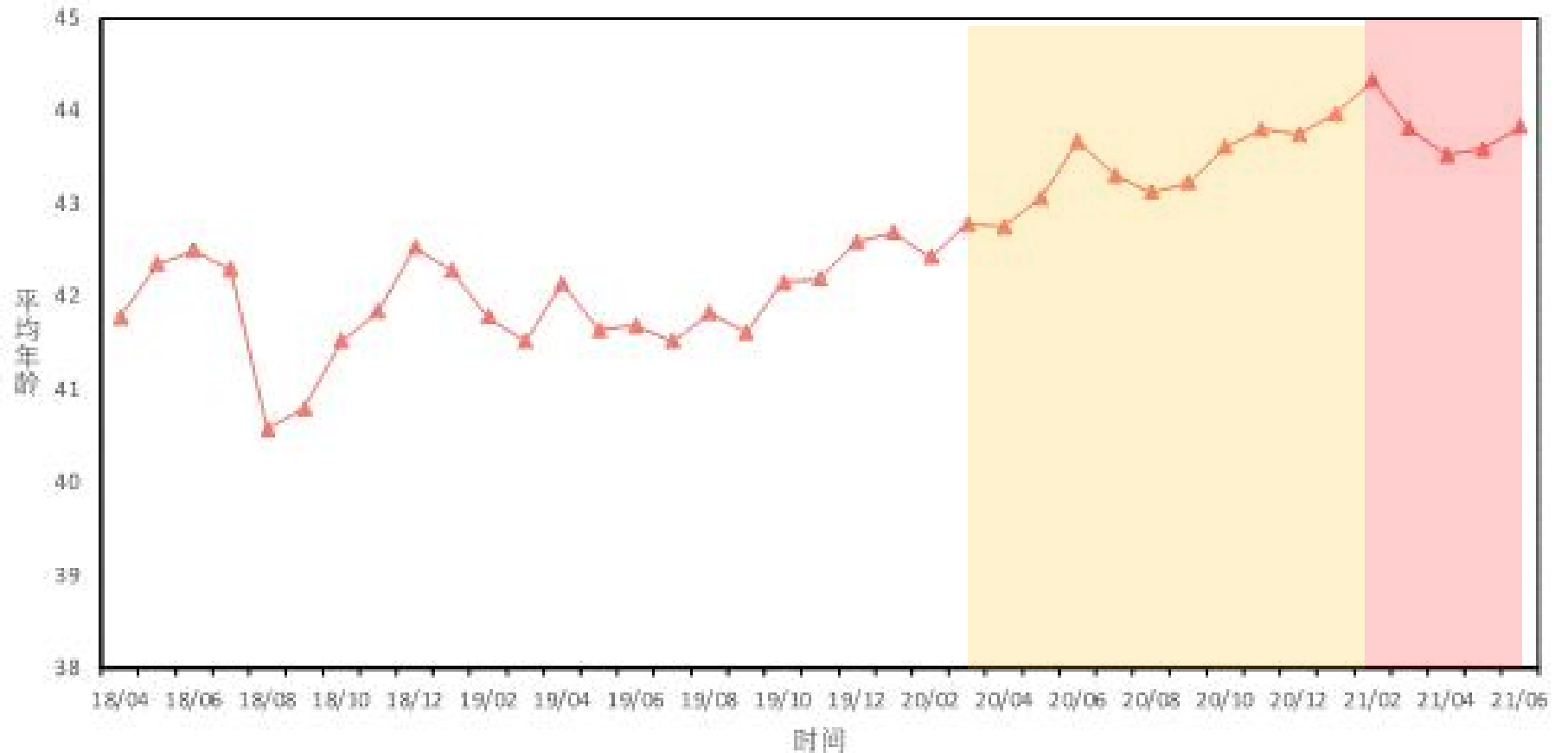
医疗费用分类型变化趋势



2.3.2 各风险指标时间变动趋势探究

- 平均参保年龄在3年间波动上升，在2021年后增长趋势明显，比2018年初提高近2岁，这其中既有年龄自然增长，保险公司为老消费者续保的因素，也与后续新参保人数下降，保险产品获客能力下降有关。平均参保年龄的提升是医疗费用通胀的原因之一。

平均参保年龄变化



2.3.2 各风险指标时间变动趋势探究

- 我们在上文将赔付通胀时期划分为稳增期和突增期，下表概括了两个时期通胀的具体归因，包括各个风险指标的变化趋势、对医疗费用通胀的影响程度以及预计未来变化趋势。

风险指标的变化趋势与通胀影响

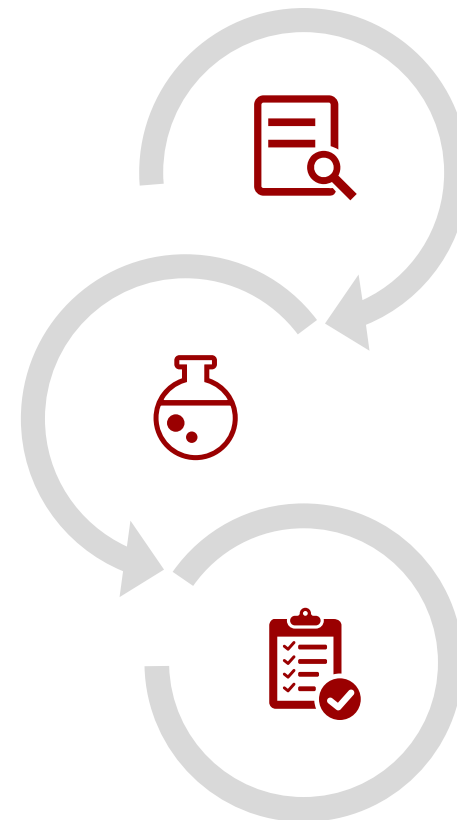
占比	稳增期 20/03~21/01	突增期 21/02~21/06	影响百分比 (稳增期)	影响百分比 (突增期)	预计未来趋势 及通胀影响
性别	女性上升	-	0.45%	-	平稳
社保类型	-	有社保下降	-	4.86%	有社保上升，削弱
重大疾病	-	有重大疾病上升	-	0.67%	重大疾病下降，削弱
医院等级	其他医院下降、二级医院上升	一级医院上升	1.27%	0.22%	其他医院上升，削弱
赔付类型	意外、一般医疗保险金下降、疾病上升	-	5.69%	-	重大疾病下降，削弱
给付类型	通融给付下降	-	0.88%	-	通融给付上升，削弱
地区	东北地区上升		5.91%	6.04%	平稳
原始总费用	上升		85.6%	87.6%	上升，增强
赔付次数	-		-	-	-
参保年龄	上升		0.59%	0.61%	上升，增强

PART 03 ▶

基于机器学习的百万医疗险赔付预测

3.1 机器学习模型构建

- 上文的分析显示，医疗通胀的产生可以归因于经济社会变化以及消费群体理赔特点的转变。持续上升的理赔费用成为产品运营和定价的一大挑战，此时通过大数据和人工智能手段预测未来赔付水平就是公司控制赔付成本的重要手段。多种机器学习模型都可以完成预测客户赔付概率和赔付金额的工作，比如经典的Logistic、SVM、决策树、随机森林等分类器，或者利用神经网络进行特征提取，并最后通过设置合适的激活函数进行预测
- 基于本组数据的特征，我们进行了对如果出险，后续赔付特征的预测，采用XGBoost集成学习算法，帮助精细刻画客户群体的整体风险特征。预测个人理赔信息能够给未来赔付成本控制、到期是否续保以及续保费率定价提供更多的决策依据。结合如果出险后的特征，可以更为精细地刻画客户的整体风险特征，无论在定价、核保、赔付，还是在早期推动精准营销都有着重要意义



3.1 机器学习模型构建

- 对每个预测指标设置损失函数计算预测误差以评价预测效果，每个指标的损失函数如下表：

预测指标损失函数

预测指标	损失函数
出险日期	预测出险日期与真实出险日期天数差距的绝对值
赔付类型	预测准确率
给付类型	预测准确率
总花费	案均差距
赔付金额	案均差距

3.1 机器学习模型构建

- 预测步骤



样本分成训练集和测试集各占50%，设定测试指标，包括性别、年龄、投保日期、地域、是否有社保5个基础指标，以及出险日期、赔付类型、给付类型、总花费和赔付金额5个预测指标

使用5个基础指标预测个体的出险日期、赔付类型、给付类型和总花费

使用5个基础指标和第一步预测得到的4个预测指标总共9个预测指标，预测赔付金额

通过定义的损失函数计算模型的误差，在训练集中对每个指标选择最合适的回溯期T作为模型的训练结果，最后在测试集中使用该流程对每个指标的当期进行预测，针对测试集中的结果评价模型的测试误差

3.2 预测结果

- 如下表，赔付金额在T=370的情况下，与真实情况的误差仅为2534元，预测表现良好

基于损失函数的预测结果评价

回溯期	10	90	170	250	330	370	测试集
出险日期*	318	305	288	268	248	233	233
赔付类型	64.50%	55.73%	55.78%	53.32%	51.54%	49.24%	63.23%
给付类型	69.64%	69.50%	67.57%	69.73%	70.93%	71.11%	67.82%
总花费	7507.88	3362.20	3043.19	3199.55	2487.12	2406.10	2315.49
赔付金额	4844.17	3270.75	3165.22	3049.69	2694.15	2664.15 (18.73%)	2533.99 (18.05%)

*这里展示的是出险日期在每个回溯期T下，样本50%分位数的绝对天数差距

PART 04 ▶

结论与建议

The background of the slide features a traditional Chinese pagoda with multiple tiers and a curved roof, situated on a hillside. The image is faded and serves as a decorative backdrop for the text.

- 本文根据HLW公司的主力百万医疗保险2018-2021年的赔付数据，系统探究了医疗通胀的原因，提出了一种基于机器学习的赔付金额预测模型。
- 在医疗通胀原因方面，研究首先指出医疗通胀可以归因为两个方面：（1）宏观方面，经济社会变动推高了总体医疗通胀；（2）微观风险指标可以左右实际赔付支出。对于宏观方面，物价水平上升带动医疗费用上涨，医疗费用的波动又传递到保险公司的赔付金额上。人口老龄化加速，预期寿命的增长，使得医疗支出压力上升，保险公司赔付盈利难度加大。除此之外，医疗进步与投入强度增大也拉高了医疗费用，使得门诊病人与住院病人的人均医疗费用均呈现了大幅上升。微观方面，研究发现投保人所处地区、性别、是否具有社会保险、就医的医院等级、出险疾病类型、赔付类型、给付类型和患病所花费用等均会对最终的保险公司赔付金额产生影响。

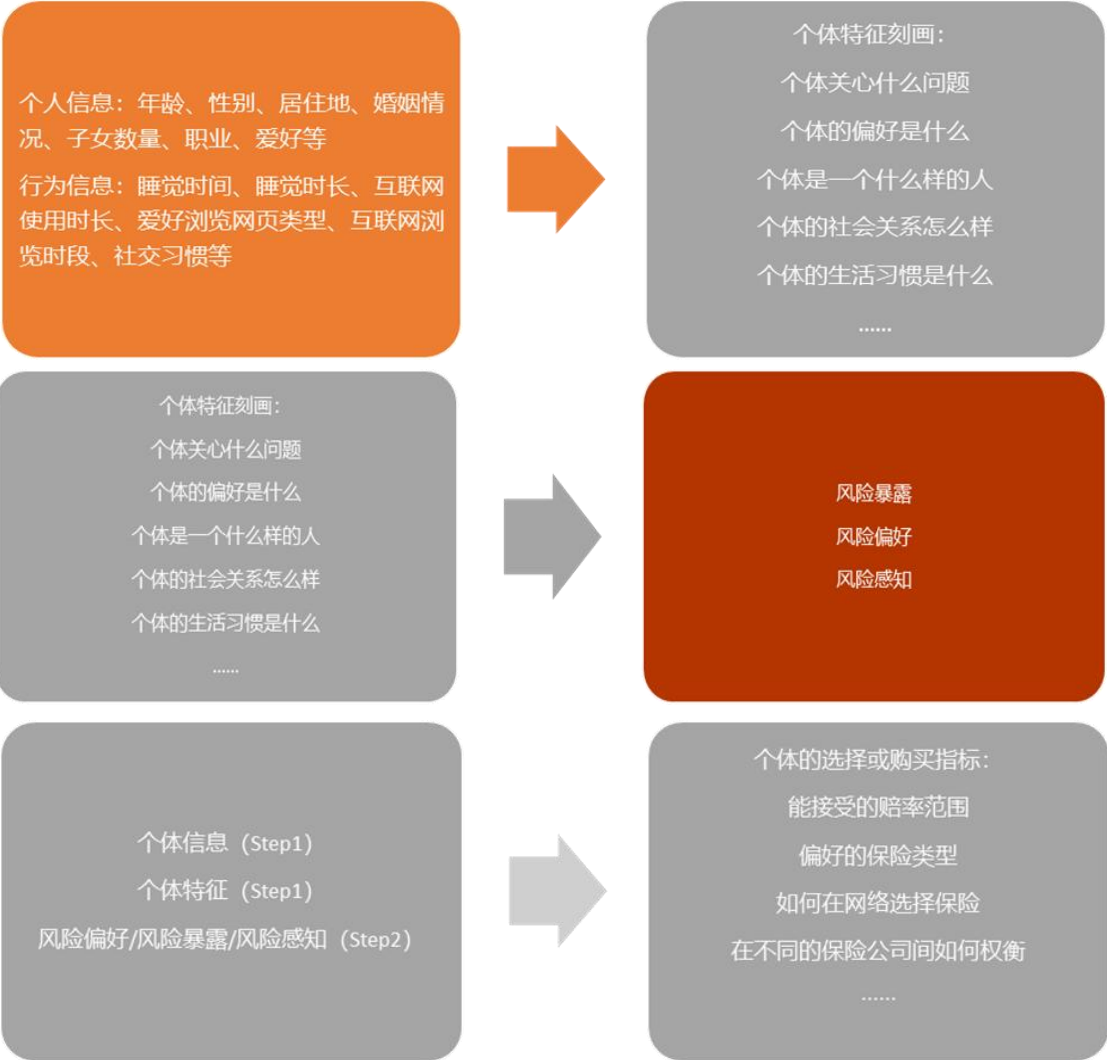


- 基于上述研究，本文提出了一种基于机器学习的赔付金额预测方法，通过每天对过去一定时期的样本回溯训练模型，并代入新投保人当天的指标值来实现对该投保人如果出险的赔付金额预测。模型使用XGBoost集成学习模型，先通过个人的性别、年龄、投保时间、所处地区、是否具备社保来预测该客户的出险时间、总花费、保险公司的赔付类型、给付类型，再结合预测的四个指标，来预测该客户一旦出险的赔付金额。研究先在随机抓取50%组成的训练集中确定每个预测模型的参数值，之后将参数值输入测试集，得到了预测偏误仅为约18%的良好表现。该模型对保险公司的运营具备指导意义，保险公司可以使用该模型结合客户出现的概率估计来更精细地评估客户风险



4.2 使用大数据对客户风险画像

- 大数据和人工智能技术的另一个应用是提供更为定制化的百万医疗险产品。目前市面上在售的百万医疗险产品保障都较为同质化，在医疗保障内容上基本一致，产品之间的差异大多体现在消费者难以感知的保额上限以及特别罕见的重疾种类上，产品竞争力的最直接体现还是在于保费价格的高低，因此保险公司容易陷入价格战的低质量竞争泥潭，在理赔费用上涨的背景下，进一步挤压了保险公司的利润空间。目前大部分医疗保险的保费定价仍然依赖于性别、年龄两个基本信息，从我们得到的理赔数据来看，对客户的其他信息利用并不充分，也无法根据客户的风险偏好定制更个人化的保险产品。对此，运用人工智能对个人风险偏好和消费习惯画像，结合模块化的产品设计方法，保险公司能够给消费者提供更为个性化的健康风险服务和产品体验



4.2 使用大数据对客户风险画像

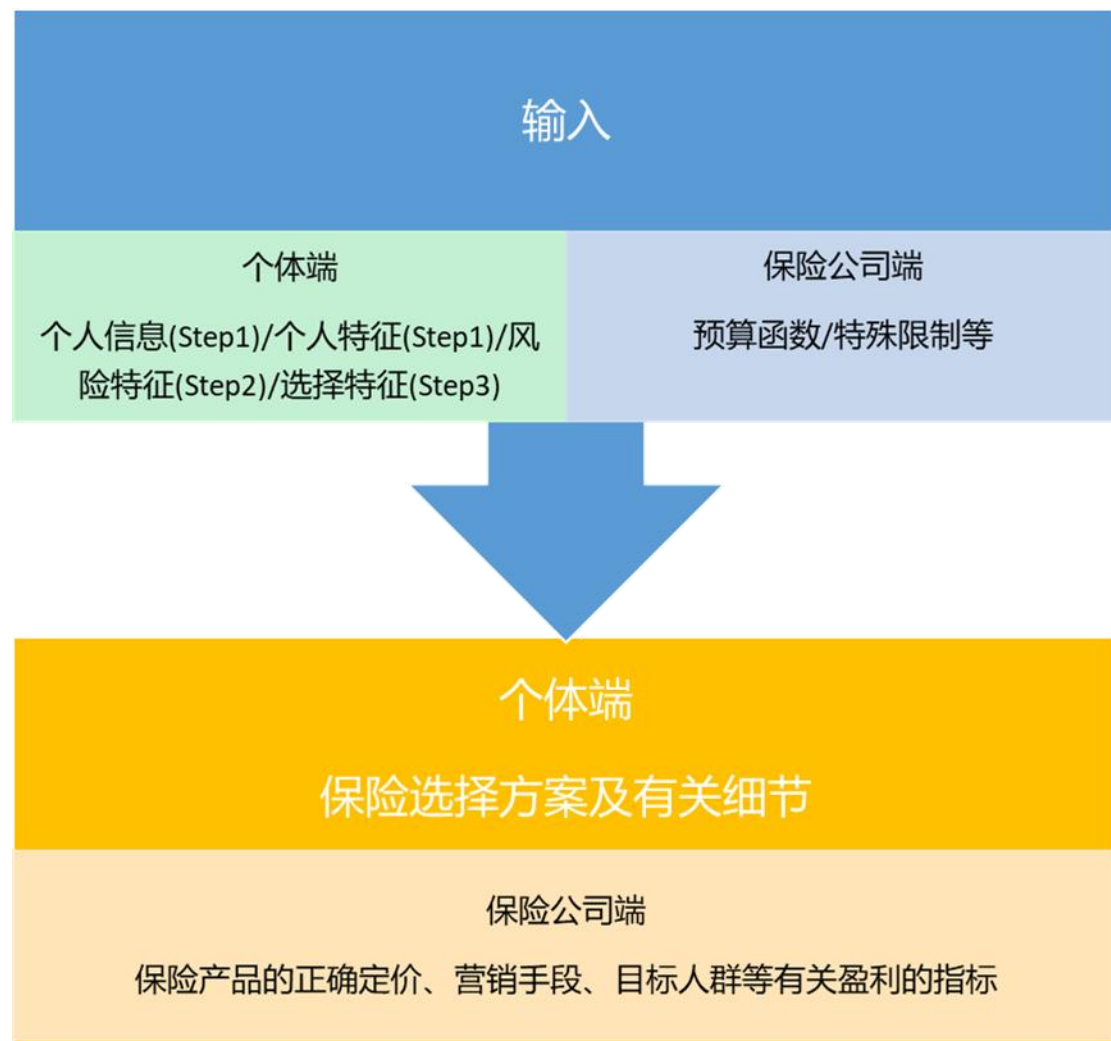
思想自由 兼容并包

- 三个在售百万医疗险产品对比

产品名称		尊享e生2022	e享护-医享无忧	平安e生保2022
保险公司		众安保险	太平洋健康	平安健康
基本规则	保障时间	1年	20年	1年
	承保年龄	0-70岁	0-65岁	0-65岁
保障内容	一般保额	300万	200万	200万
	一般免赔额	1万	1万	5千/1万可选
	重疾保额	600万/100种	400万/120种	200万/120种
	重疾免赔额	0 (60岁以上1万)	1万	5千/1万可选
	住院前后门急诊	前后30天	前后30天	前后30天
	门诊手术	有	有	有
	重疾津贴	1万 住院: 100元/天(可选)	1万	/
	质子重离子	100%/600万	100%/400万	100%/400万
	外购药	有	有	有
增值服务	费用垫付	有	有	/
	赴外医疗	可选	/	可选
	就医绿通	有	有	有
续保条件	满期未停售	需审核	需审核	需审核
	停售续保新品	需审核	需审核	需审核
保费测算		首年保费 (有社保)		
	30岁	293	238	350

4.2 使用大数据对客户风险画像

- 新技术从核保、承保到理赔全链条都能发挥重要作用。在数据来源方面，需要医疗机构、互联网平台甚至其他保险公司联通，构建统一跨平台的数据库，推动医疗健康行业的信息互通。在核保的环节，通过收集用户以往的医疗记录结合个人基本信息和财务数据，构建详细完整的个人医疗健康画像，在核保环节运用人工智能技术向客户输出合适的产品；在出险前的日常管理阶段，通过风险画像给予客户相应的健康风险提示，配合可穿戴设备或医疗器械，打通医疗机构和家庭医生的沟通渠道，监测用户的健康状况，及时给予健康信息建议和提醒；在后续的理赔环节，可以采用自动化理赔、大数据反欺诈等手段，减少理赔阶段的人力成本和沟通成本，同时对异常就医行为、异常医疗支出、异常赔付行为进行监测，及时发现保险欺诈行为，切实降低理赔成本。



4.3 与“惠民保”差异化竞争

思想自由 兼容并包

- 在产品定位上，惠民保和百万医疗险同属中低端的健康保险，相较于百万医疗险，惠民保突出了政府宣传背书、保费更低、允许非标体和老年群体投保三个优点，满足了广大人民群众对医疗保险负担性和信任度的需求。尽管百万医疗险的赔付范围和比例都要比惠民保更高，但在保障范围大部分重叠和保费更低的情况下，大部分投保人认为两种产品可以互相替代，因此他们往往会选择购买价格更为低廉的惠民保，使得百万医疗险的市场地位受到冲击

		本产品	京惠保2022
投保年龄		30天-65周岁	年龄不限
核保方式		有健康告知	无健康告知
续保		出险人群可以续保 但不保证续保	政府鼓励持续经营
保障范围	住院医疗	不区分医保目录内外	住院医保目录内费用限自付一 住院医疗自费医疗费用
	特殊门诊	√	×
	门诊手术	√	×
	住院前后门急诊	住院前后7天	×
	恶性肿瘤特药	可选	√ 共60种
	质子重离子	可选	√
	免赔额	一般为10000元 重疾为0元	住院费用20000元 其余保障为0元
	报销比例	100%	根据保障项目 50%、90%、100%不等
价格		相对较高 根据年龄不同存在差异	79元
增值服务		就医绿通、视频问诊、多学科会诊	特药服务、健康咨询、慢病管理、健康体检等七大类24种

4.3 与“惠民保”差异化竞争

思想自由 兼容并包

- 为了应对惠民保的强势竞争，百万医疗险需要明确其中端健康险的定位，在产品设计上突出与惠民保的差异化。惠民保专注于满足基本的医疗保障需求，而百万医疗险可以满足相对而言更先进，更定制化的医疗需求

报销外购药

部分医院由于医保资金限制等原因无法在院内开药，需要患者自行在外购药，报销这部分无法被医保报销的医药费能够减轻患者无法在医院开药被迫外购药物的经济负担

垫付医疗费用

在医疗机构提出支付要求时给患者垫付医疗费用，减轻患者短期的支付压力

地区和服务差异化

针对该地区常见多发的癌症或慢性病，增加治疗和药物报销目录和范围，对接大型公立医院或私立医院专家科室，提供更专业更高端的诊疗服务。产品设计和宣传上突出更全面更高端的保险保障，以满足消费者获取全方位医疗保障和提升医疗服务体验的需求

PART 05 ▶

创新点和应用价值



5.1 创新点

- **医疗通胀是我国目前存在的一大现象，百万医疗保险产品的赔付情况也受到了医疗通胀明显的影响，是保险公司关注的重要课题。现存的描述更多在于对医疗通胀问题进行总体上的分析，但是关于其对于百万医疗保险赔付的影响以及对应的归因分析比较稀少**

01

本文利用宏观分析结果和微观风险指标模型，系统性对医疗通胀引起的百万医疗保险赔付成本上升进行了归因：经济社会变动推高总体医疗通胀、微观风险指标左右实际赔付支出

02

基于归因指标对医疗通胀的影响，本文使用机器学习算法构建了采用相关指标数据对赔付的预测模型，预测表现优异

03

本文进一步指出了未来可研究的方向：基于更多种数据对客户的风险画像和未来赔付预测，有助于保险公司进一步降本增效

5.1 研究或应用价值

- 本文的研究结果对于保险公司有着一定的实践意义

01

保险公司可以利用研究得到的归因体系，系统评估当下及未来的赔付风险情况，提早做出应对，降低赔付成本，调整商业布局

02

保险公司可以利用预测模型，更加精细地对个体的风险水平进行评估，制定个性化的保险方案以更好匹配用户需求

03

保险公司可以进一步对研究中指出的未来方向进行探究，利用大数据分析技术，提取个体特征，进行更加精准的风险评估，升级营销手段，在行业竞争中创造优势

感谢聆听

T H A N K S F O R Y O U R L I S T E N I N G

2022.06