医疗知识图谱的构建及应用

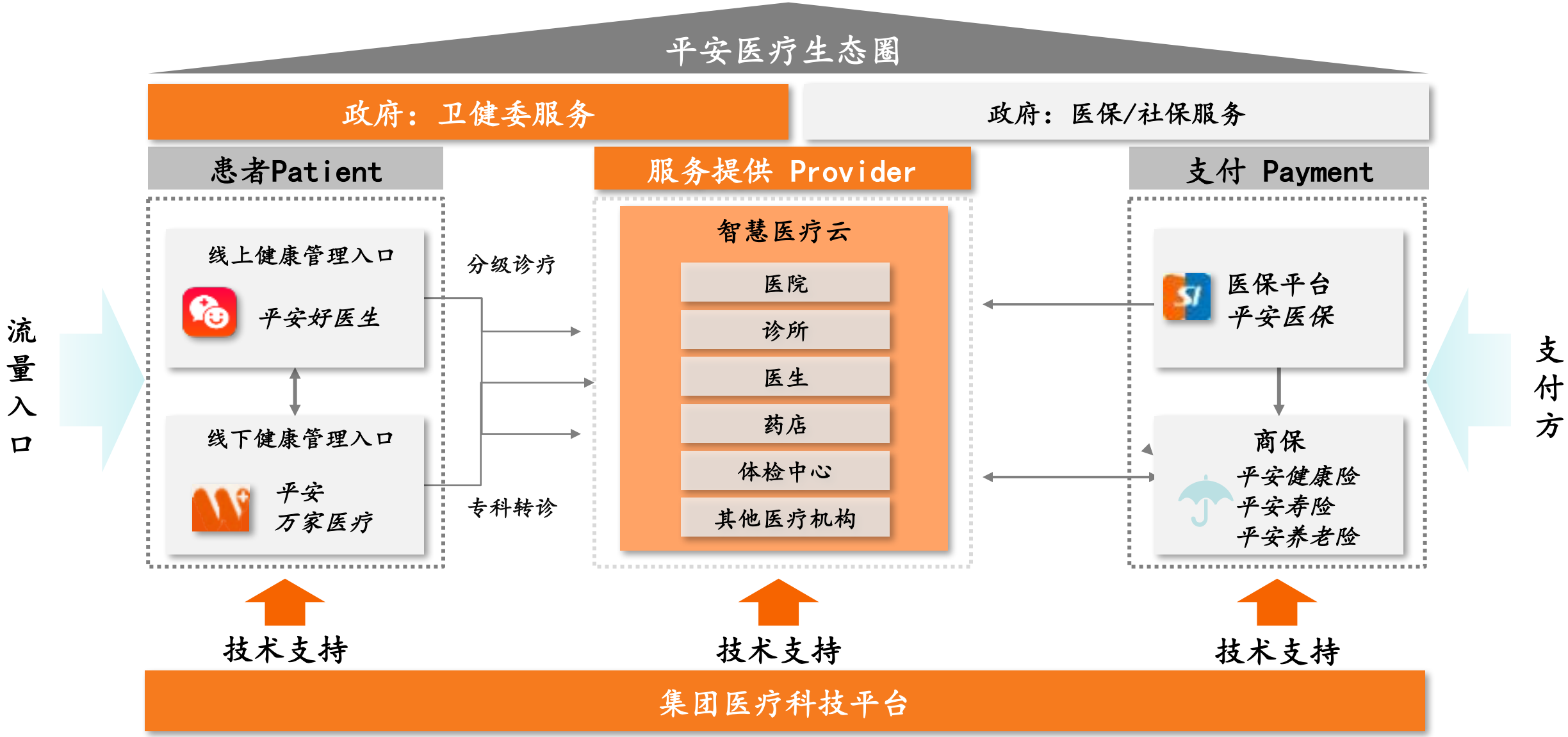
中国平安
保险·银行·投资

平安医疗科技
PING AN HEALTHCARE TECHNOLOGY

倪渊，平安医疗科技医疗文本处理负责人

2018年12月8日

平安集团PPP模式医疗生态圈



集团医疗科技平台

A4 6大智能服务

预防	诊前	诊中		诊后/管理	质控
智能疾病预测	智能预诊/分诊	智能影像筛查/辅助诊断	智能疾病辅助诊疗	智能随访追踪	智能质量控制

A3 医疗知识图谱

医疗知识图谱（覆盖60万医学概念，530万医学关系，1000万医学证据）

A2 5大信息库

疾病库　治疗库　药品库　个人健康库　医疗机构和医生库

A1 N个数据源

院前			院中			院后		
健康体检	健康档案		院内信息系统	第三方影像	队列研究&临床试验	院外随访	商保核赔	
在线问诊	个人监测	流行病	电子病历	基因测序		医保核赔	患者问题	药品购买

A2 5大医疗信息库

疾病库

- 疾病及疾病相关属性和关系
- 常见疾病知识库
- 多个城市地方病种库
- 不同版本ICD编码
- 疾病语义概念
- 疾病同义词库
- 症状同义词库

治疗库

- 治疗模式
- 精准人群
- 检查检验信息
- 手术信息
- 临床路径
- 经典病例

药品库

- 药品信息
- 各省市药品目录
- 药品中标价格
- 药物不良反应
- 药物疗效
- 临床试验

个人健康库

- 疾病风险因子
- 患者常见问答
- 疾病知识问答库
- 科普文章
- 生活方式

医疗机构&医生库

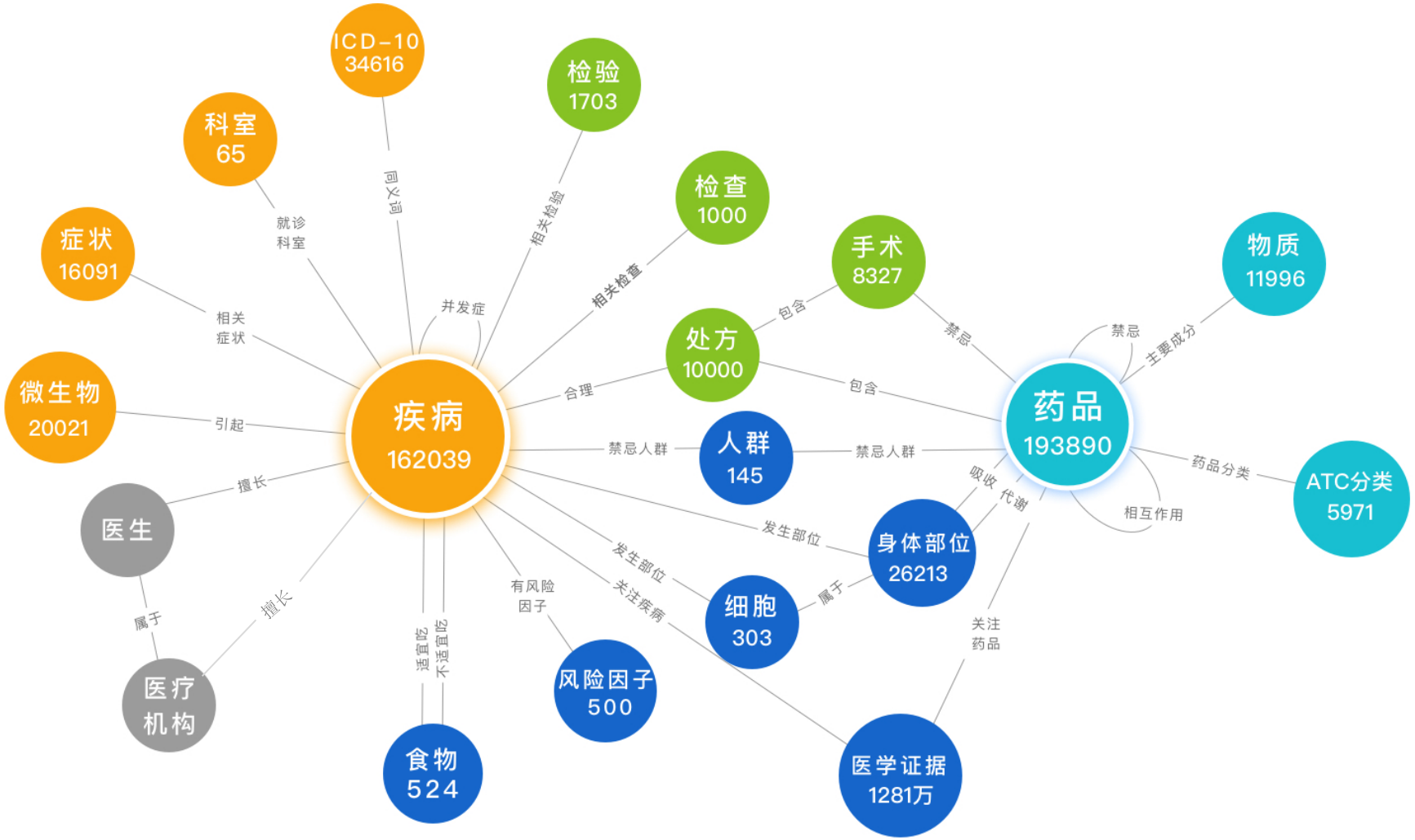
- 医疗机构&医生名称
- 医疗机构级别
- 主要科室
- 专家人数
- 住院床位
- 医生简介
- 医生资质
- 医生特长

A3 平安医疗知识图谱：涵盖50类医学概念，191类医学关系

中国医学科学院 北京协和医学院
医学信息研究所 图书馆
Institute of Medical Information / Medical Library, CAMS & PUMC

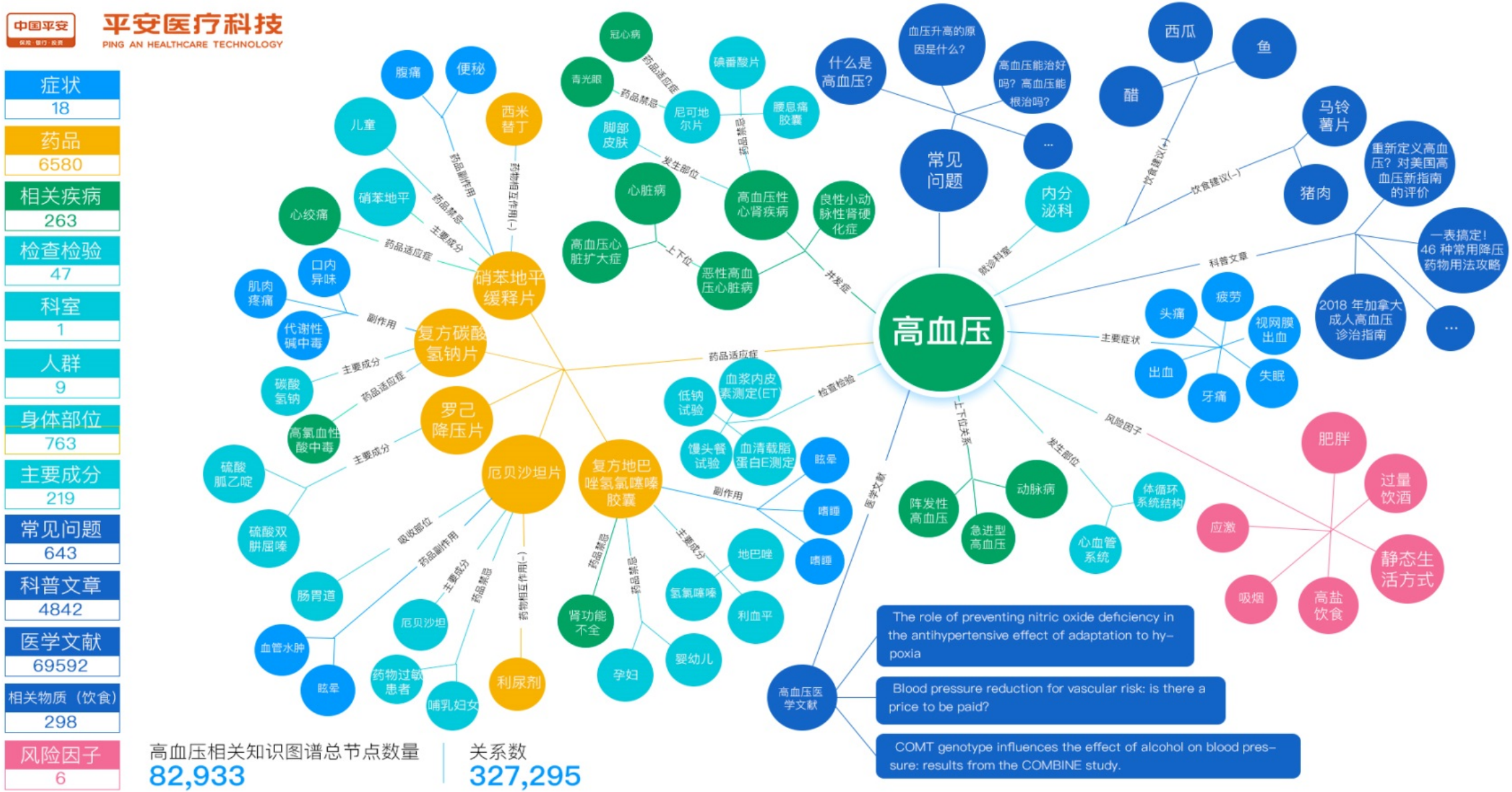
OMAHA

知识图谱：以**结构化的形式**描述客观世界中的**概念**及其**关系**



- 50类医学概念
- 191种医学关系
- 100种医学属性
- 60万医学术语
- 530万医学关系
- 1000万医学证据

A3 知识图谱：高血压知识图谱示例



平安医疗知识图谱-整体架构

融合医学核心概念以及医学临床证据

智能服务

- 医疗知识图谱查询导览及可视化
- 基于自然语言的知识图谱交互
- 基于知识图谱的医患教育
- 基于知识图谱的结构化和标准化
- 基于知识图谱的决策支持

数据平台

- 爬虫
- MongoDB
- 图数据库
  - Cassandra
  - HBase
  - BerkeleyDB
  - ElasticSearch
- 自然语言处理平台
- 机器学习平台

图谱构建

- Schema管理
  - 知识图谱schema管理工具
- 数据结构化，槽填充
  - 实体识别
  - 属性识别
  - 实体链接
  - 关系抽取
  - 模板定义和管理
- 数据驱动
  - 基于数据挖掘的相关性
- 多图谱融合
  - 实体对齐
  - 对齐审核工具
- 图谱质量评价
  - 图谱冲突解决工具
  - 图谱一致性校验
- 图谱规则集
  - 规则校验工具
  - 推理规则

数据层

医学核心概念

OMAHA 开放医疗与健康联盟

中国医学科学院 北京协和医学院 医学信息研究所 图书馆 Institute of Medical Information / Medical Library, CAMS & PUMC

- 疾病知识库
- 检查检验知识库
- 症状知识库
- 药品知识库
- 身体部位知识库
- 手术知识库

+

医学临床证据

经典病例

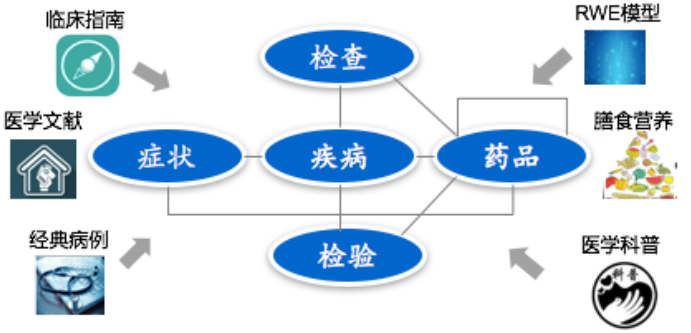
RWE模型

临床指南

医学科普

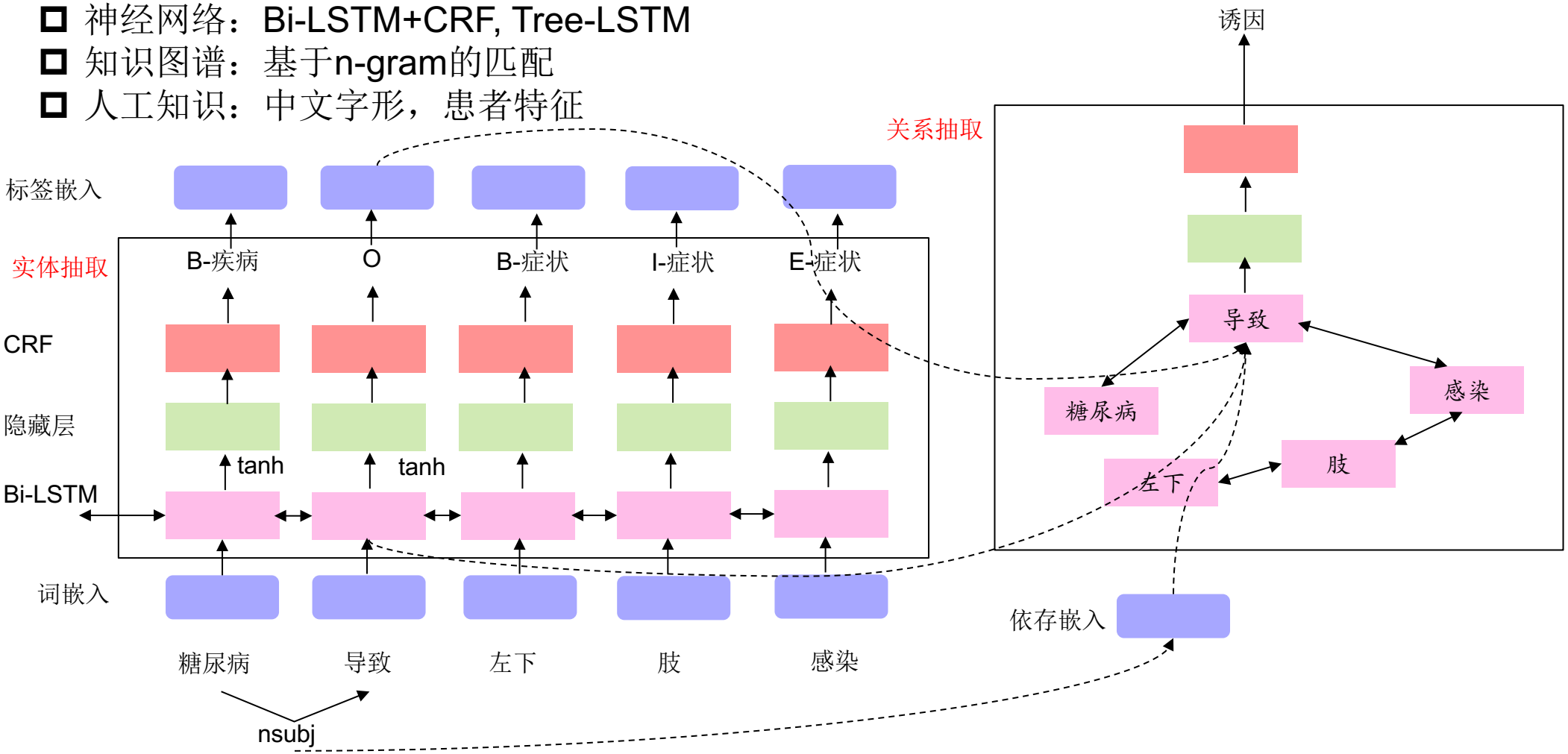
=

医学知识+证据图谱



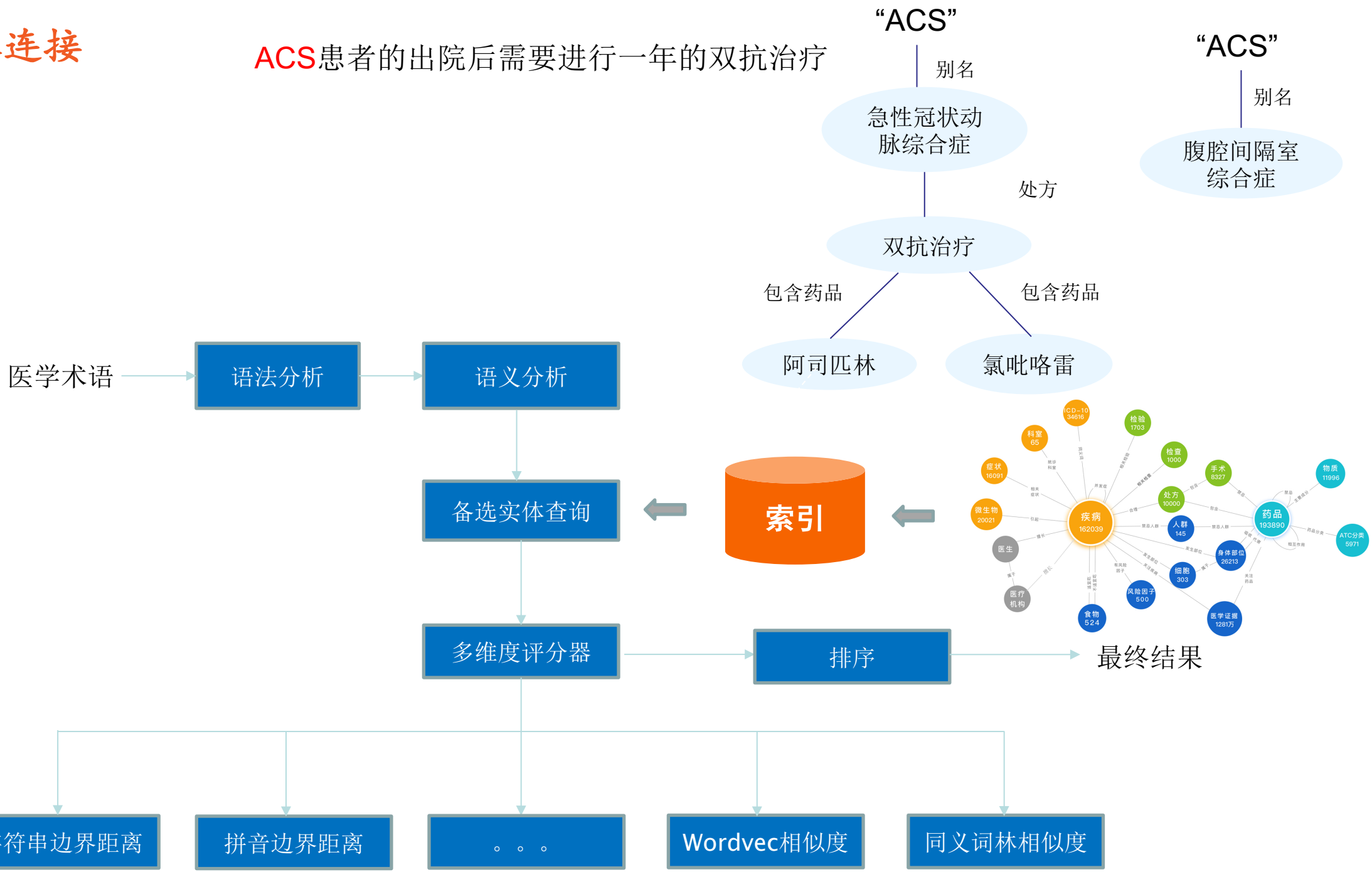
实体识别&关系抽取

- 基于深度学习的端到端的命名实体识别以及关系抽取，利用多任务方式同时提高命名实体识别以及关系抽取的精度
- 神经网络+知识图谱+人工知识
  - 神经网络：Bi-LSTM+CRF, Tree-LSTM
  - 知识图谱：基于n-gram的匹配
  - 人工知识：中文字形，患者特征

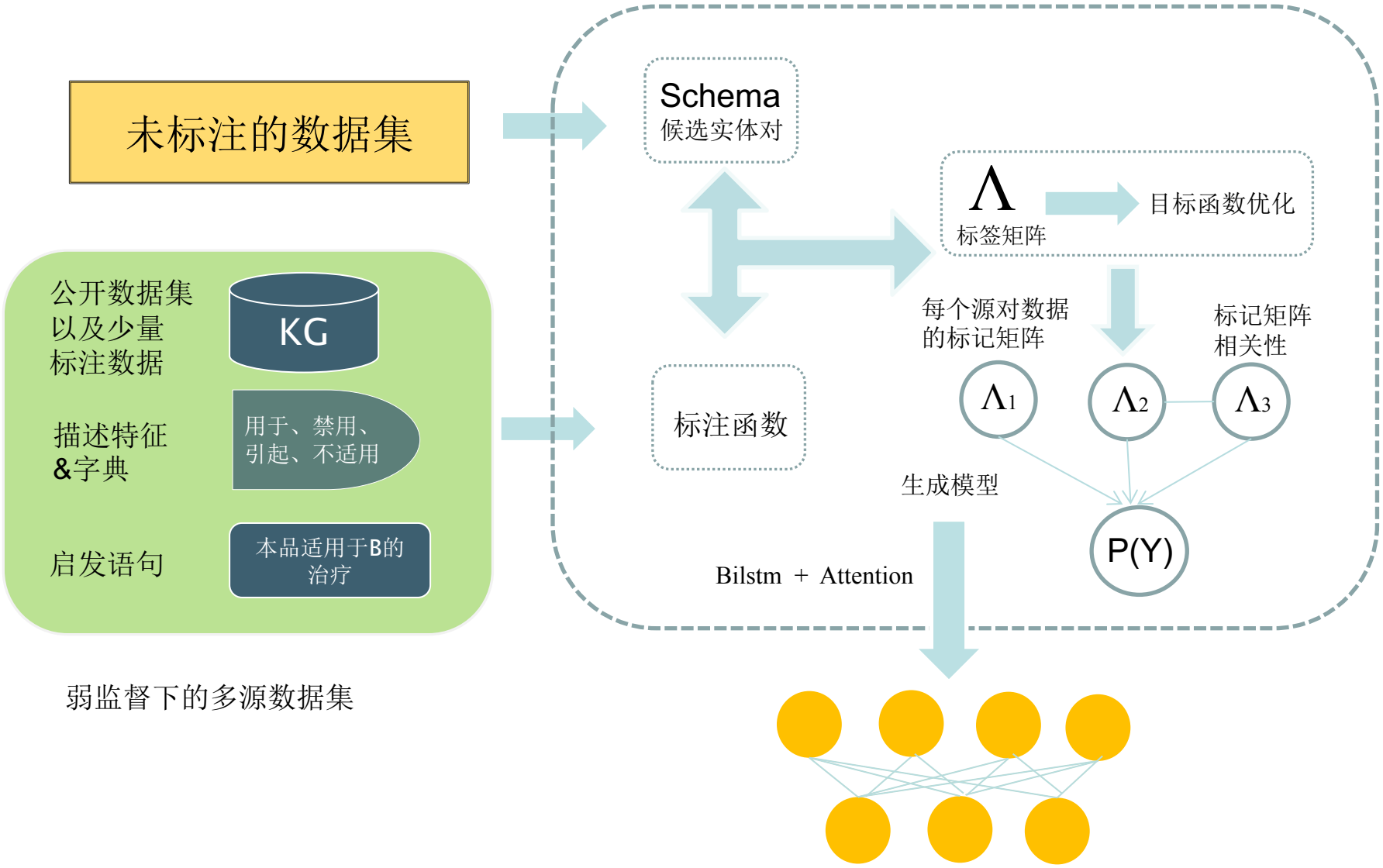


实体连接

ACS患者的出院后需要进行一年的双抗治疗



弱监督的信息抽取



弱监督下的多源数据集

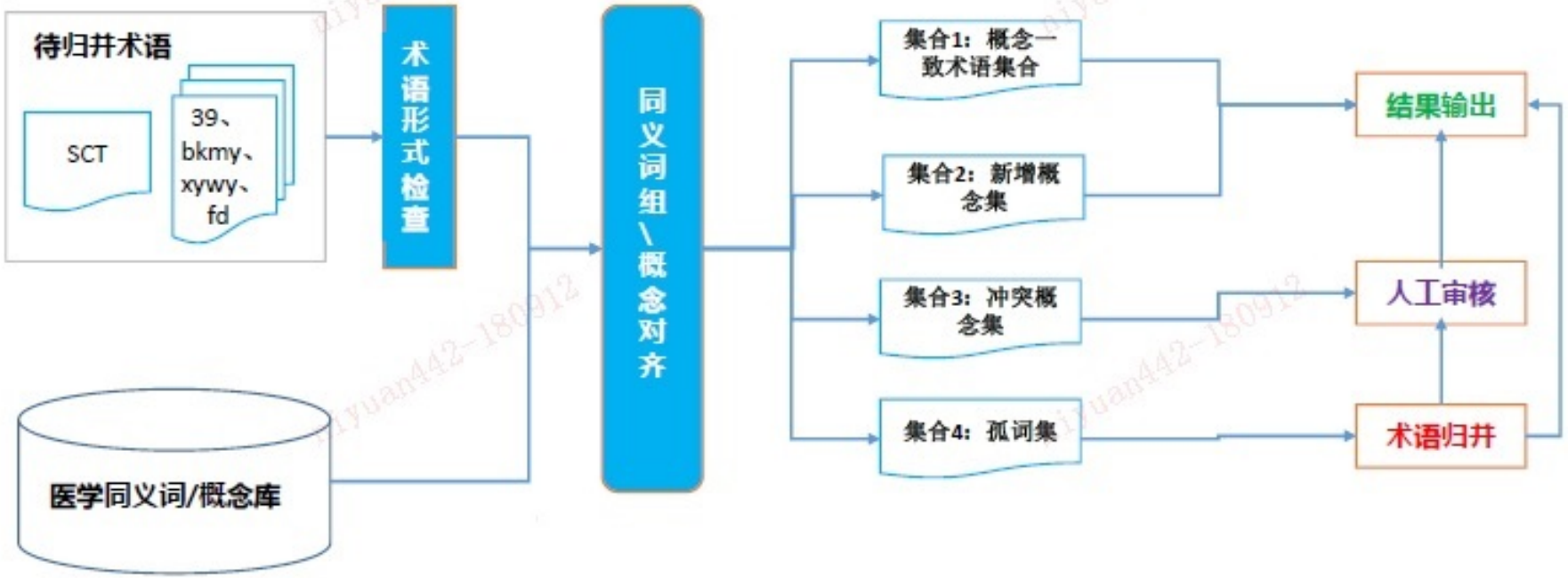
图谱融合

医疗领域存在多种形式的同义词

- 不同翻译名称
  - 阿司匹林
  - 阿士匹灵
  - 阿斯匹林
- 拼写错误
  - 天竺癇
  - 天竹黄
- 别名
  - 小檗碱
  - 黄连素片

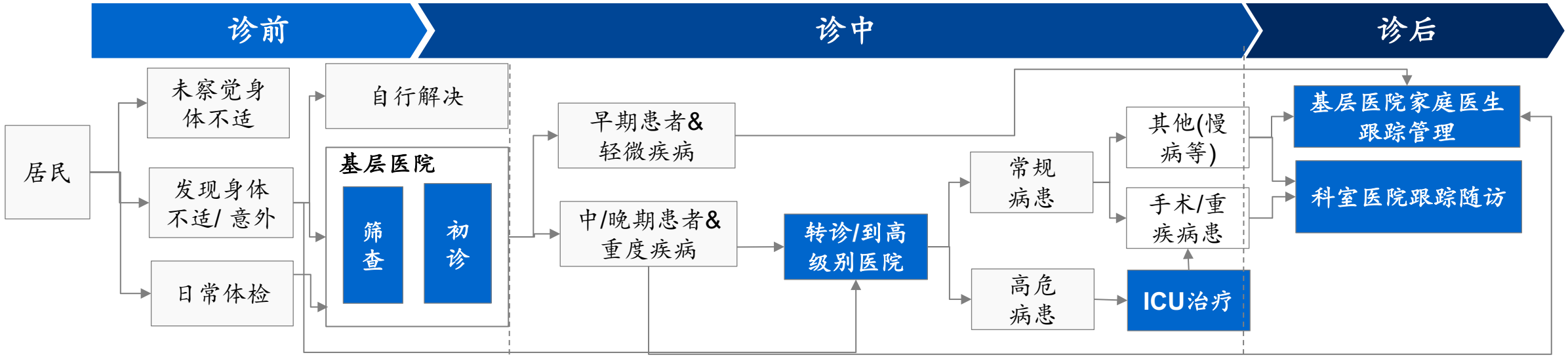
阿托伐他汀相关药物

阿托伐他汀分散片
阿托伐他汀片
山乐汀阿托伐他汀钙片(社区)
阿托伐他汀片(山乐汀)
氨氯地平/阿托伐他汀钙片
氨氯地平阿托伐他汀片(多达一)
阿托伐他汀片(立普妥)
阿托伐他汀片(阿乐)
阿托伐他汀钙分散片
阿托伐他汀钙片
阿托伐他汀钙片(医) 20mg/tab(**)
阿托伐他汀钙片(合资)
阿托伐他汀钙片(山乐汀)
阿托伐他汀钙片(山德士)
阿托伐他汀钙片(立普妥)
阿托伐他汀钙片(立普妥)H
阿托伐他汀钙片(立普妥20mg
阿托伐他汀钙片(阿乐)
阿托伐他汀钙片(限)
阿托伐他汀钙片_(立普妥)
阿托伐他汀钙胶囊
阿托伐他汀钙胶囊(国产)
阿托伐他汀钙胶囊(限)



医联体/医院在诊前、诊中、诊後存在诸多痛点

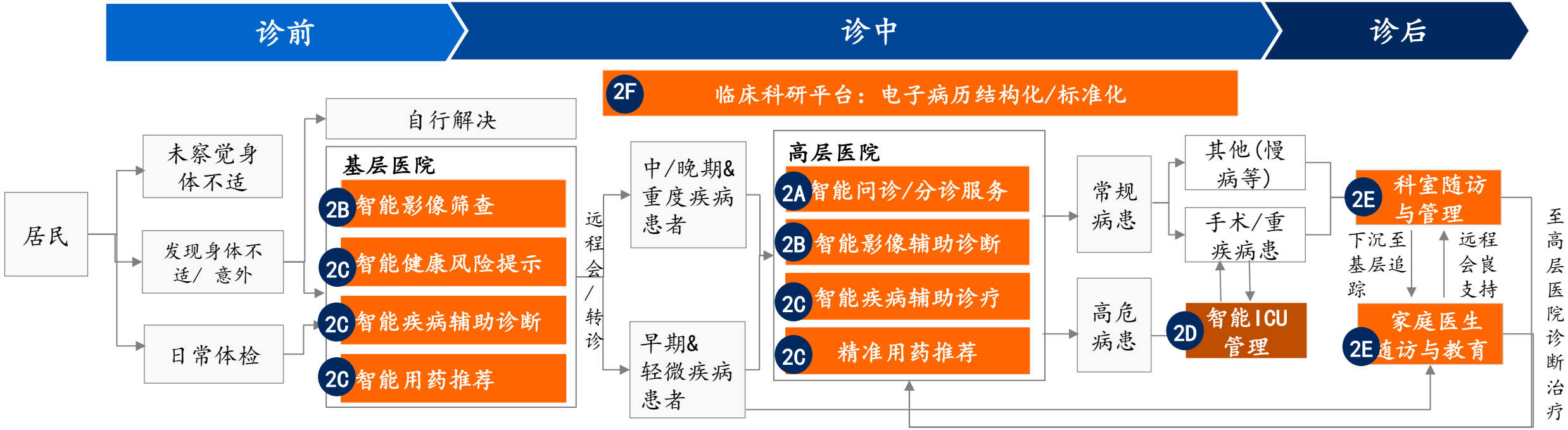
关键环节



		诊前	诊中	诊后
主要痛点	医院	• **基层诊疗水平低, 误漏诊高(~40%):** 370万基层卫生人员中, 本科以上学历仅占9%，中级以上职称仅占12% • **基层缺乏专科与病理医生, 设备利用率低** (如CT设备利用率<40%)	• **放射、病理医生负担大, 误漏诊高:** 日均接待200-300病人, 每个患者读片时间不到2分钟 • **临床医生时间未有效利用:** 20-50%以上时间於文字报告入录; 小病、常见病、慢性病患者诊疗量占大头 • **重症/ICU病患量大, 管理难:** 平均每重症患者1/6个医生, 工作压力巨大，难以有效管理病患 • **院内系统众多，数据孤岛难以整合信息**	• **缺乏有效诊后管理/随访手段:** 无有效数据统计、工作量繁重缺乏主动管理辖区病患环节
	病患	• **三级医院挂号难:** 病患量大, 挂错号比例高 • **疾病发现晚:** 卫教信息落後, 缺乏渠道与医生沟通	• **病患挂错诊号比例高(10-30%[1]):** 许多诊疗资源浪费 • **医院等待/排队时间长:** 挂号、候诊、影像断、交费、拿药占>90%时间; 医生医生问诊时间短	• **慢症患者没有定期至医院复诊,** 导致晚期发现, 情况加剧

1 2015年首都医科大学附属北京朝阳医院急诊科主任郭树彬统计; 前几大误挂包括皮肤科误挂外科(~31%), 外科误挂内科(~30%), 口腔科误挂外科(~12%), 外科误挂妇产科(~11%), 中医误挂外科(~10%)

智慧医院CDSS解决方案全景图



2A 智能分诊 /导诊

- 全病种(候诊大厅与药剂科)诊前数据收集&机器问答智能分诊
- 带来高效且精准的智能就医体验

2B 多模态智能影像辅助诊疗

- 提供~35种基于影像的辅助诊疗
- 包含智能诊报告&远程放射会诊等模块，提升诊疗效率，降低误诊率

2C 智能疾病辅助诊疗与治疗推荐

- ~30种常见疾病智风险提示、能辅助诊疗与用药建议，规范临床诊疗和用药标准
- 精细管理医疗资源，提升诊疗准确性

2D 智能ICU管理

- 死亡时间预测和诊疗资源预测、管理等智能模块
- 有效提高医生诊疗品质与效率

2E 智能患者教育/随访

- 全病种基于自然语言理解的智能随访和诊后患教
- 协助基层医院提升诊后管理与随访效率

2F 智能临床科研平台

- 利用深度学习技术，自动抽取电子病历中的关键信息
- 基于医疗知识图谱的医学概念标准化

2A 2C 诊前诊中辅助决策

带有概率及权重的知识图谱
节点粗细：节点的重要性/频度；边粗细：相关性概率

诊前(分诊)

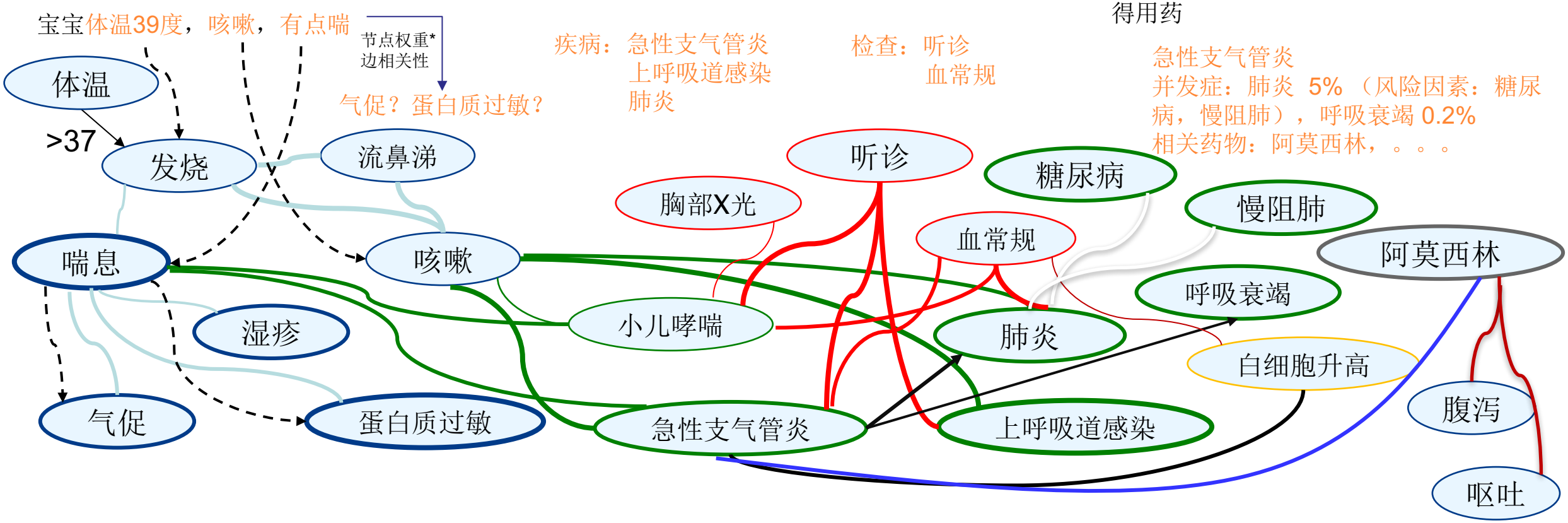
基于症状相关性以及重要度的推荐，完善主诉

诊中(诊断辅助）

基于患者主诉，选择相关疾病，并按照节点权重及相关性进行排序。同时进一步给出排序的推荐检查

诊中(治疗辅助)

检查检验结果的解释。根据医生诊断，提示可能得并发症及预后。提示可能得用药



2C 个性化精准诊疗-辅助诊断

数据准备

美国20年救护车使用调查公共数据集，包含210万次就诊记录

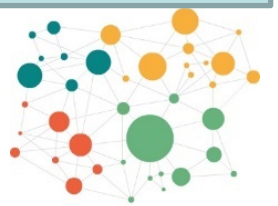
全球最大的医学文献数据库

7百万篇医学文献摘要中提取出的症状，诊断以及共现关系

斯坦福大学临床数据仓库

120万病人跨度19年的2000万次就诊记录中抽取的疾病（症状和诊断），药物，设备和手术信息

医学知识

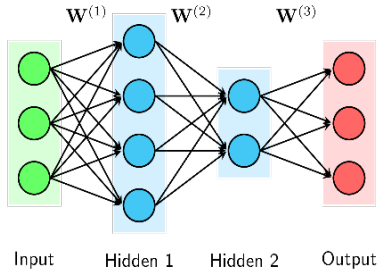


医学知识图谱和疾病的临床指南。包括1.2万个症状概念，14万个疾病概念，以及相互关系

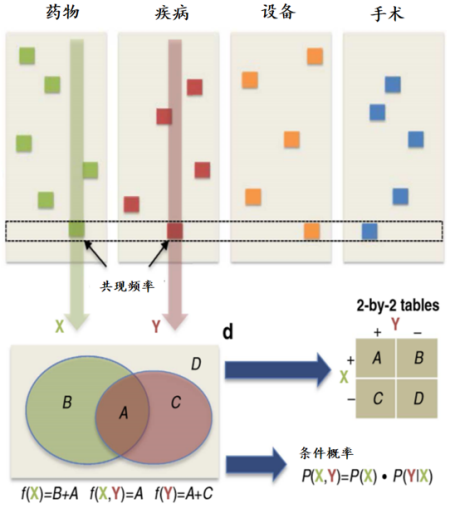
数据建模方法

构建深度学习模型

- 通过医学知识图谱及中的疾病和症状关系构建**基础网络**
- 利用公共数据进行增量学习， 学习优化目标考虑先验知识 – 在损失函数中加入**知识对应的罚项**，提高模型精度。



构建贝叶斯网络



- 应用 PubMed 和 Stanford 数据构建**概率模型**，支持更多种类的疾病
- 通过**模型融合技术**生成最后的诊断列表

模型应用- 辅助诊断

研究价值和进展

- **价值：**利用患者症状、体征等信息为医生推荐疑似诊断，减少误诊、漏诊
- **进展：**概率模型覆盖500+种疾病，全科常见30+种疾病的诊断模型的准确率为95%*
- **合作：**北京大学第一医院全科

2C 个性化精准诊疗-辅助治疗

数据准备

临床指南

- 收集整理了20+国家临床指南和专家共识
- 包括特定疾病的诊断、**治疗**和预防的方法和路径。

知识图谱

- 药品： 17万种
- 药品和疾病关系：35万条
- 药物相互作用：21万条

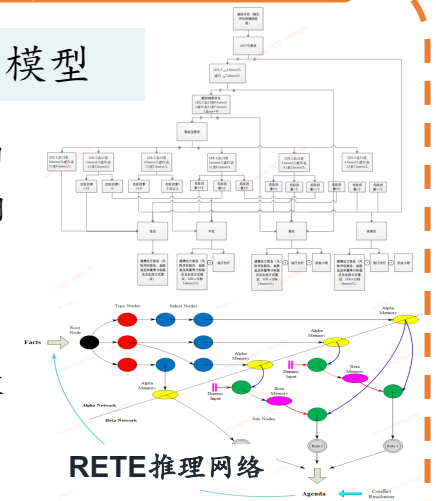
临床数据

- **重庆卫健委电子病历数据：**包括35种常见疾病的就诊115万次。
- **理赔数据：**包括35种常见病就诊1700万次。
- 每次**就诊**包括患者基本信息，诊断，检验检查信息，疾病史，用药史和处方信息。

建模方法

基于知识的治疗模型

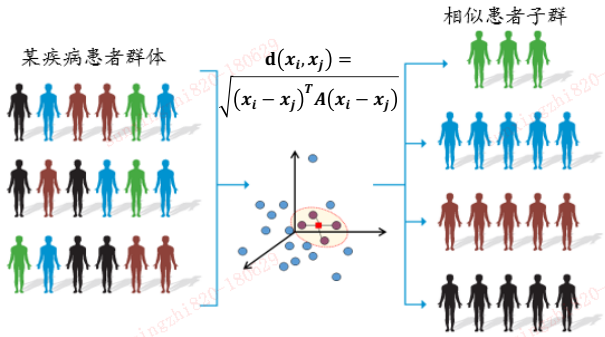
- **知识表示：**将临床指南表示为**决策树**，进而翻译成可执行的**规则**。
- **知识推理：**采用**推理引擎**将规则应用到患者数据，生成治疗方案。



RETE推理网络

基于数据的治疗模型

- 通过**关联规则分析**的发现常见治疗模式。
- 通过**精准分群技术**，找出在临床上相似患者的人群。
- 在相似人群中的进行治疗模式的**个性化推荐**。
- 采用**模型融合技术**把推荐结果和知识模型结果整合



模型应用- 药物推荐

- 根据患者的信息给出**处方推荐**

用药推荐

用药方案	药品通用名	药品商品名	规格	用法	剂量	频率	价格	循证
血压管理	ACEI/ARB	卡托普利	25mg	口服	2片	Q.d.	78元	
血糖管理	二甲双胍	盐酸二甲双胍片	0.5g	口服	1片	Tid.	28元	
血脂管理	阿托伐他丁	阿托伐他丁钙片(立普妥)	20mg	口服	2片	Q.d.	62元	
血小板管理	阿司匹林	阿司匹林肠溶片	0.1g	口服	2片	Q.d.	16元	

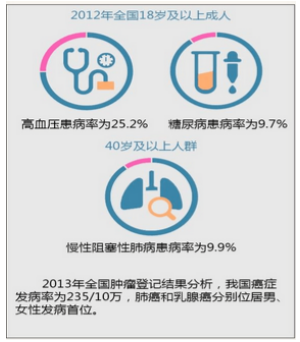
- 提供**推荐依据：**

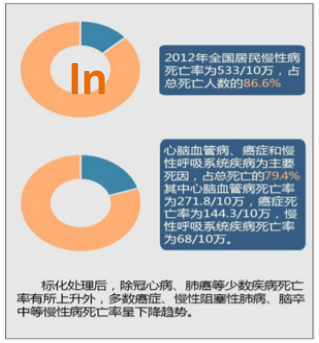
研究价值和进展

- **价值：**规范医生治疗，提升患者满意度
- **进展：**完成3种疾病（高血压，糖尿病，房颤）的依据知识的治疗模型，正在开发基于数据的30+种疾病的治疗推荐模型
- **合作：**上海中山医院全科、赛诺菲

2E 疾病健康问答场景

3亿高血压患者， **1亿**糖尿病患者，
5000万慢阻肺

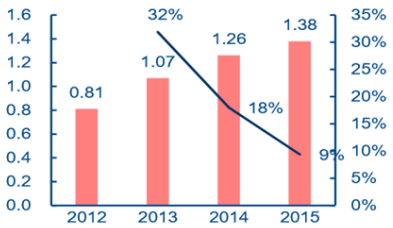




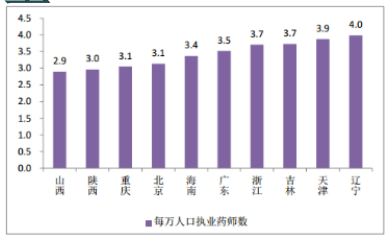
社区卫生服务中心

药店 PHARMACY

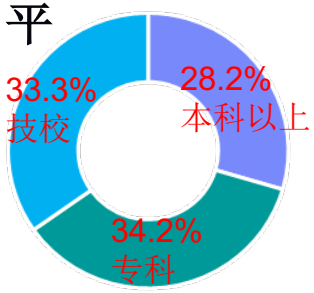
每1万人的社区医生数量



每1万人的药师数量



药师的教育水平



患者在慢病管理当中会有各种疑问

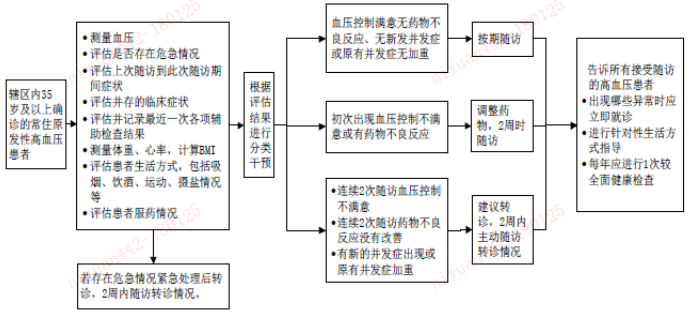
平安好医生　好大夫在线 www.haodf.com　上亿注册用户

寻医问药网 闻康 XYWY.COM　春雨医生　百万级日问诊量

大量健康教育相关问题，可以从医疗知识库当中找到答案

糖尿病患者应该如何饮食？
立普妥和络活喜可以一起吃吗？
空腹血糖8.6，有问题吗？

卫计委要求家庭医生要定期对慢性病随访



随访内容包括

症状：头痛，头晕，胸闷等
体征：血压，体重,BMI，心率等
生活方式：吸烟，饮酒，运动等
用药：种类，用法用量

"超过5亿人"的家庭医生去哪了_网易新闻

数据丰满,现实骨感,5亿人的家庭医生都在哪儿?-中青在线

5亿人有家庭医生？医生:连病人家门在哪都不知道|家庭医...-新浪新闻

2E 疾病健康问答-问题相似度匹配

数据驱动的FAQ问题集准备

患者问题库
（4000万+）

是否FAQ → 问题向量表示 → 问题聚类 → FAQ挑选 → 检索模型 → 语义匹配 → 训练集生成

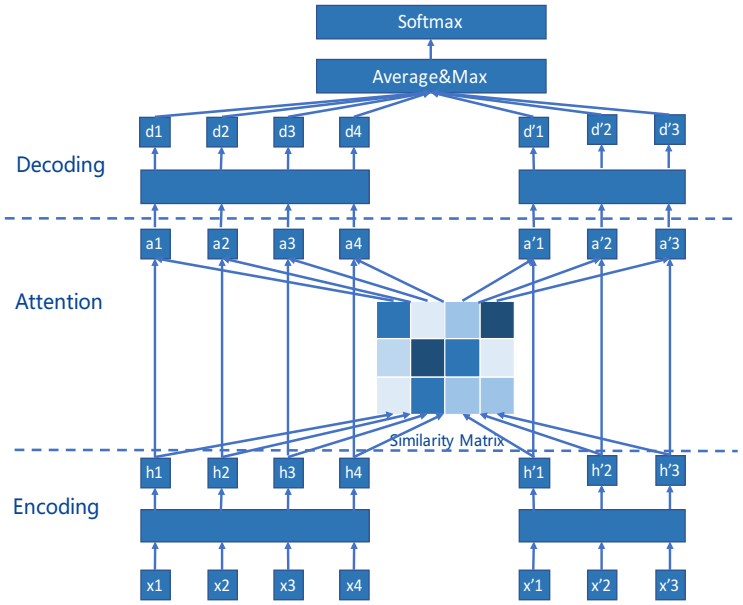
300w+问题对

FAQ库

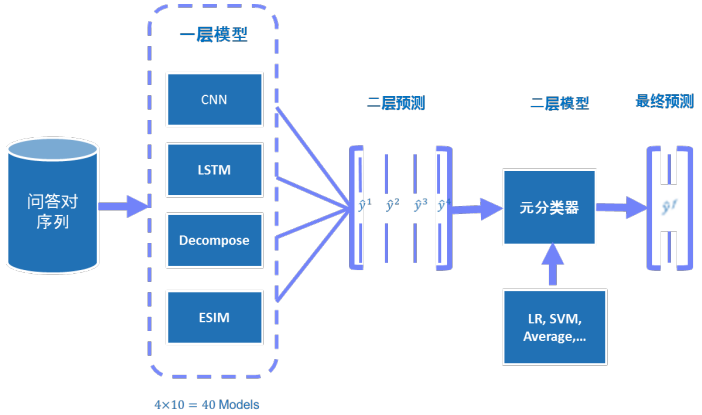
2000+FAQ覆盖糖尿病，高血压，房颤

2E 疾病健康问答-问题相似度匹配

基于深度注意力网络的文本匹配



集成学习技术



病种	问题库	FAQ	问题组	训练集	精度
糖尿病	170, 000	90, 000	700	340万对	93%
高血压	110, 000	80, 000	707	110万对	95%
房颤	10, 000	5，000	200	10万对	92%
卒中	60, 000	20, 000	350	35万对	92%
慢阻肺	20, 000	10, 000	437	14万对	91%

2E 药品助手-基于知识图谱的问答

模板

<药品>一天应该吃几次？

备选关系:
适应症，用法用量，
禁忌，。。

词向量
矩阵

模板
编码

关系向量

关系-模
板匹配

匹配
组合

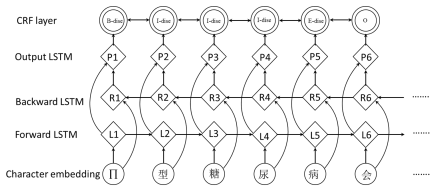
Max Avg

MLP

用法用量

阿托伐他丁一天应该吃几次？

命名实体识别
biLSTM + CRF



阿托伐他丁一天应该吃几次？

实体链接

阿托伐他汀钙片 | 阿托伐他汀钙分散片 | 阿托伐他汀钙胶囊 | 阿托伐他汀胶囊

多轮对话的实体消岐

片剂 | 胶囊

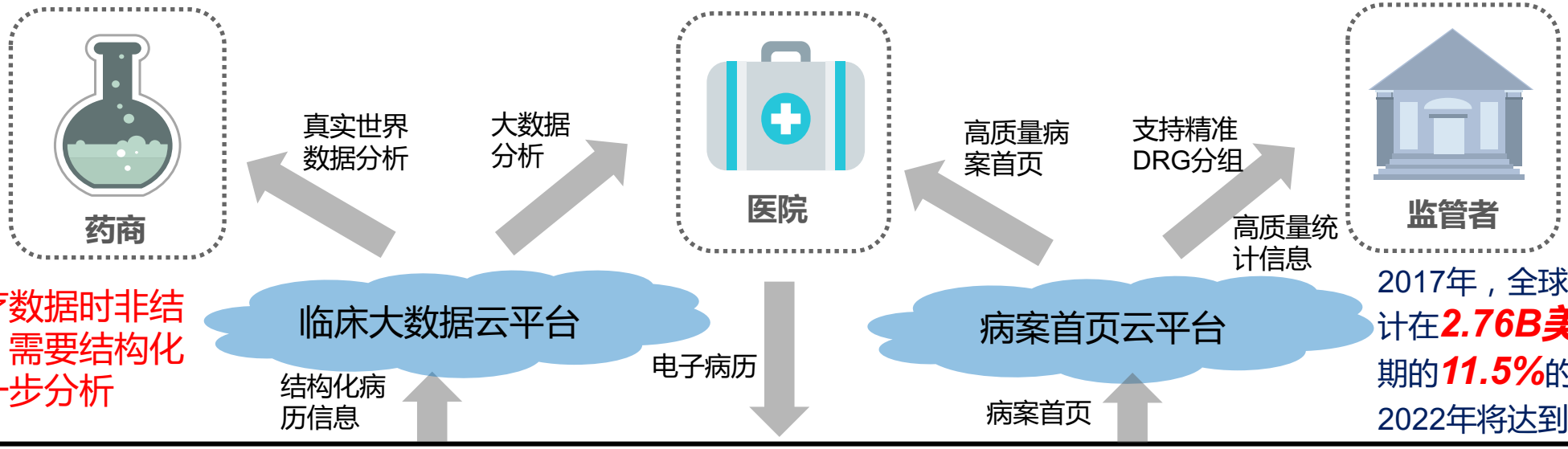
立普妥 | 阿乐

立普妥(10mg) | 立普妥(20mg) | 立普妥(40mg)

立普妥（10mg)常用的起始剂量为10 mg每日一次…

2F 基于知识图谱的电子病历理解和智能编码

每年***70亿人次***就诊患者，产生***100亿+***电子病历数据；通过电子病历结构化以及自动编码服务，整合电子病历数据



80%的医疗数据时非结构化数据，需要结构化后才能进一步分析

2017年，全球自动编码市场估计在***2.76B美金*** 并且按照预期的***11.5%***的CAGR，到2022年将达到***4.75B 美金***[1]

患者诉发冷，出现寒战，头晕，恶心未吐。血常规：白细胞数17.83*10^9/L，。。。查体：左肾区叩击痛阳性,未触及肿块。。。考虑患者泌尿系感染合并SIRS。制定治疗方案：1停止静点头孢替唑，给予头孢哌酮舒巴坦钠2.03/日静点

自然语言理解理解

症状	有	寒战，头晕，恶心
	无	呕吐
检验	白细胞	17.83*10^9/L
体征	有	左肾区叩击痛阳性
	无	肿块
诊断		泌尿系感染合并SIRS
治疗	无	头孢替唑
	有	头孢哌酮舒巴坦钠

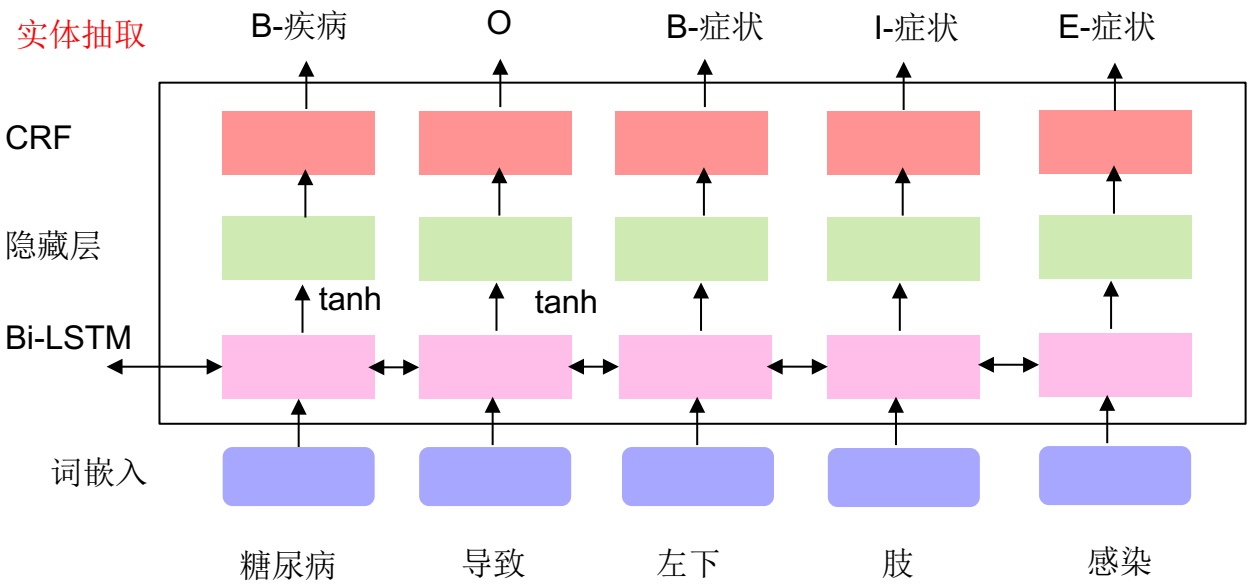
自动编码

- ICD-10疾病编码
- ICD-9-CM手术编码

[1]https://www.marketsandmarkets.com/PressReleases/computer-assisted-coding.asp

2F 信息抽取相关技术

- LSTM+CRF 模型



- 模型只能解决部分问题
  - 无恶心，呕吐，无发烧，发冷
  - 尿尿尿不出尿来3天
  - 全身浅表淋巴结未触及肿大，头颅无，五官端正
  - 今查房，症状同前，体查同前，治疗同前
  - 巴林斯基征，巴氏征，巴宾斯基征，巴宾斯基，babinski征，巴彬斯基征，babinskisign。。。

2F 电子病历结构化和标准化产品

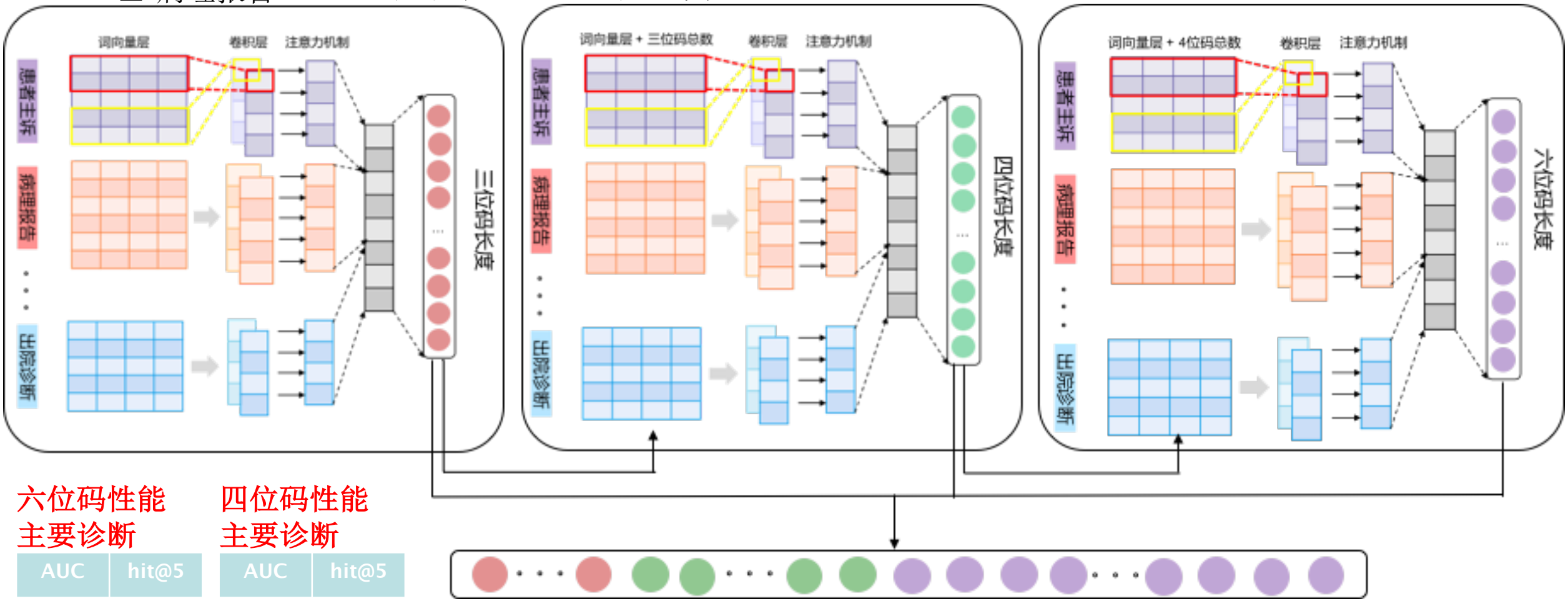
数据看板，快速浏览结构化电子病历的统计信息

自动抽取电子病历文本中的信息（例如疾病，症状，症状属性等），并利用知识图谱将信息标准化。生成标准化的结构化数据

支持对于结构化电子病历的语义查询

2F 基于机器学习的自动编码

- 考虑整体的电子病历来提取特征
  - 患者主诉
  - 病理报告
  - 就诊过程
  - 出院诊断
  - 出院医嘱
  - 患者基本信息
- 通过机器学习来自动的推荐编码，包括主要诊断，次要诊断以及手术编码
- 通过医学知识图谱来进一步提高性能



六位码性能
主要诊断

AUC	hit@5
0.99	0.89

四位码性能
主要诊断

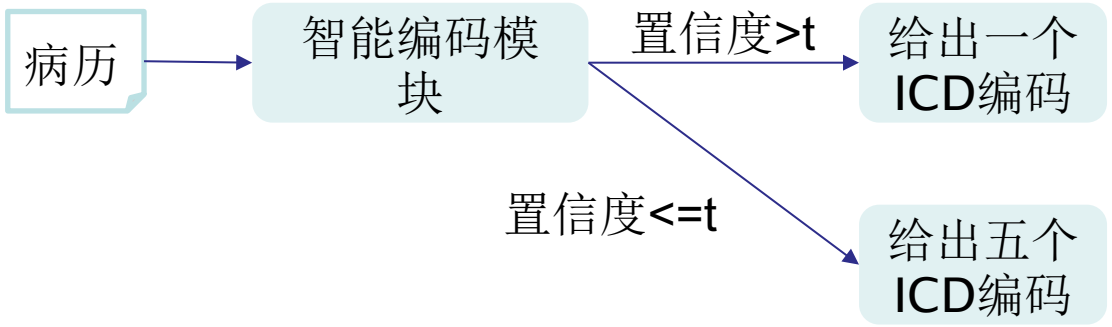
AUC	hit@5
0.99	0.92

2F 模型比较

训练数据：22万+电子病历，包含临床诊断描述，科室，患者信息等和icd编码

模型（Label）	Hit@1	Hit@5	coverage hit@1=0.95	coverage hit@1=0.90
模型1(一层六位码模型)（六位码）	0.62	0.85	2%	5%
模型2(层级模型)（六位码）	0.66	0.85	17%	28%
模型3(层级多任务模型)（六位码）	0.73	0.89	20%	32%

hit@1 = 模型只给出一个备选编码的时候就是正确编码的百分比
hit@5 = 模型给出5个备选编码的时候包含正确编码的百分比
coverage @（hit@1=0.95） 在hit@1可以达到0.95的时候，模型可以覆盖的病历百分比
coverage @（hit@1=0.90） 在hit@1可以达到0.90的时候，模型可以覆盖的病历百分比



准确度高于95%，无需人工审核

给出5个备选编码，供人工挑选

减少20%的人工审核时间

和搜索查找比较，将编码推荐范围从～30降低到5，减少30～40%的人工

2F 基于知识图谱的医学文献查询

医学文献

命名实体识别

实体链接

关系抽取

属性抽取

主题提取

情感分类

研究

P

疾病

人群

干预方式

I

治疗

比较方式

C

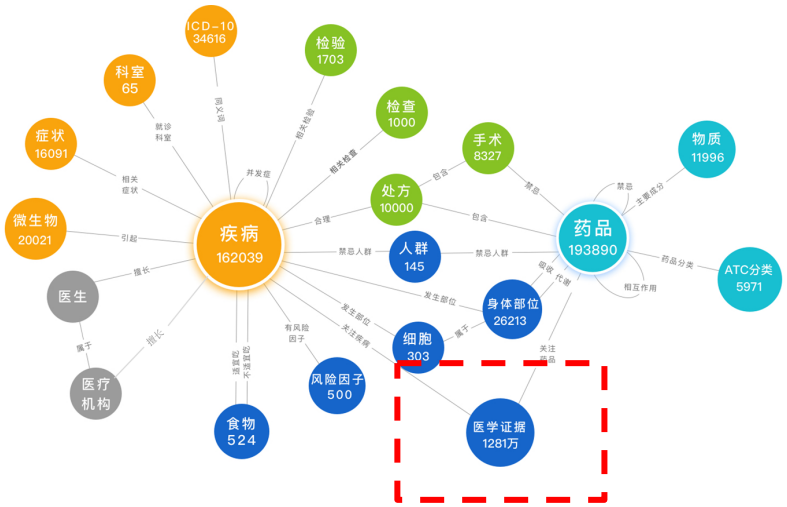
有效
无效
中立

O

结局

方法

实验方法

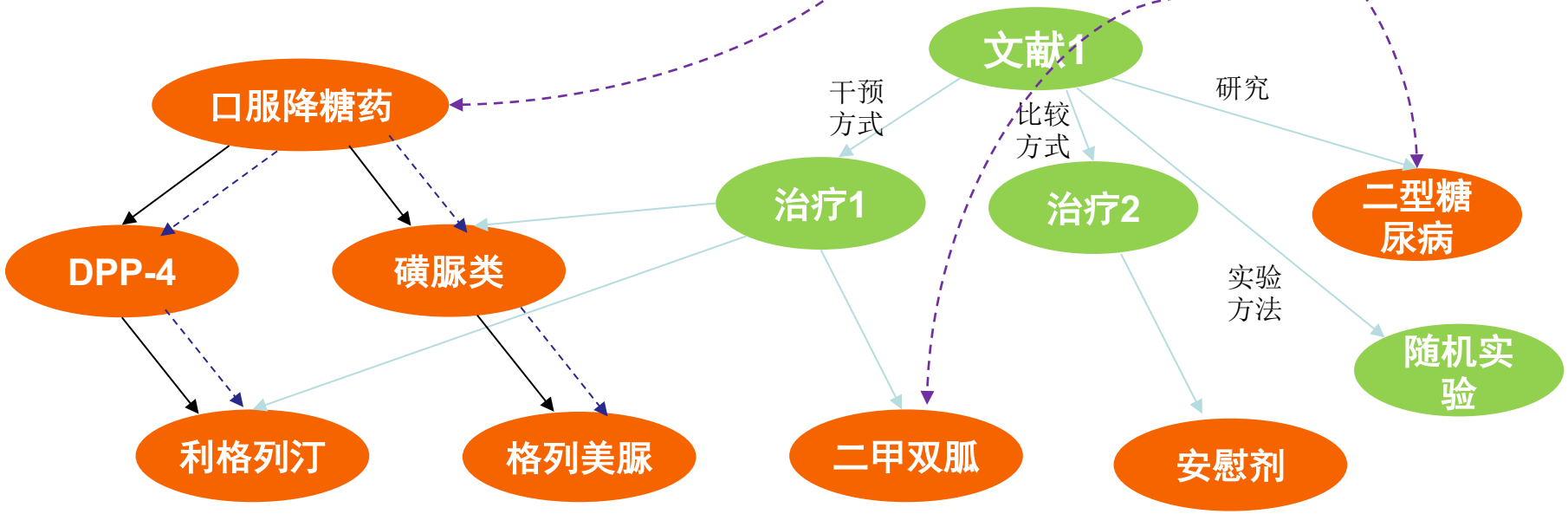


2F 基于知识图谱的医学文献查询

利格列汀联合二甲双胍及磺脲类药物在2型糖尿病患者治疗中有效性、安全性分析

摘要：目的： 研究利格列汀联合二甲双胍及磺脲类降糖药物治疗（二甲双胍及磺脲类药物血糖控制不佳的）的2型糖尿病患者有效性及安全性的相关数据。方法：48例2型糖尿病患者按2：1比例随机分为利格列汀组与安慰剂组，两组均用药24周。结果：利格列汀与二甲双胍及磺脲类降糖药物联合应用24周可以显著降低糖化血红蛋白（HbA1c）的水平（与安慰剂组比较，HbA1c较基线时下降0.62%，P<0.0001）。结论：利格列汀与二甲双胍及磺脲类降糖药物联合应用安全有效。

2型糖尿病患者，已经在服用二甲双胍，血糖控制不满意，加入哪种口服降糖药合适？



Thanks