基于中文电子病历的心血管疾病风险因素标注体系及语料库构建

苏 嘉 [1]　何 彬 [1]　吴 昊 [2]　杨锦锋 [3]　关 毅 [1]
姜京池 [1]　王焕政 [1]　于秋滨 [2]

摘　要　本文讨论了从中文电子病历中标注心血管疾病风险因素及其相关信息的问题, 提出了适应中文电子病历内容特点的心血管疾病风险因素标注体系, 构建了中文健康信息处理领域首份关于心血管疾病风险因素的标注语料库.

关键词　心血管疾病, 中文电子病历, 风险因素, 语料标注, 自然语言处理

引用格式　苏嘉, 何彬, 吴昊, 杨锦锋, 关毅, 姜京池, 王焕政, 于秋滨. 基于中文电子病历的心血管疾病风险因素标注体系及语料库构建. 自动化学报, 2019, **45**(2): 420−426

DOI　10.16383/j.aas.2018.c170206

Annotation Scheme and Corpus Construction for Cardiovascular Diseases Risk Factors From Chinese Electronic Medical Records

SU Jia[1]　HE Bin[1]　WU Hao[2]
YANG Jin-Feng[3]　GUAN Yi[1]　JIANG Jing-Chi[1]
WANG Huan-Zheng[1]　YU Qiu-Bin[2]

Abstract　In this paper, the issue of annotating cardiovascular diseases (CVDs) risk factors and the related information from Chinese electronic medical records (CEMRs) is discussed and an annotation scheme of CVDs risk factors appropriate to the content characteristics of CEMRs is put forward. Furthermore, the first annotated corpus of CVDs risk factors in the field of Chinese health information processing is constructed.

Key words　Cardiovascular diseases (CVDs), Chinese electronic medical records (CEMRs), risk factors, corpus annotation, natural language processing

Citation　Su Jia, He Bin, Wu Hao, Yang Jin-Feng, Guan Yi, Jiang Jing-Chi, Wang Huan-Zheng, Yu Qiu-Bin. Annotation scheme and corpus construction for cardiovascular diseases risk factors from Chinese electronic medical records. *Acta Automatica Sinica*, 2019, **45**(2): 420−426

心血管疾病 (Cardiovascular diseases, CVDs) 是一组发生于心脏和血管的疾病[1], 据 2015 年世界卫生组织报道, 心血管疾病已成为世界范围内的头号死因, 每年死于心血管疾病的人数多于任何其他疾病. 而在发展中的中国, 情况更为严重. 世界卫生组织明确指出, 大多数的心血管疾病可以通过控制诸如烟草使用、不健康饮食、肥胖、缺乏锻炼等风险因素而得到预防[1]. 其他的心血管疾病风险因素还包括高血压、糖尿病、血脂异常等[2]. 因此, 对这些风险因素的管理控制成了预防心血管疾病的首要任务. 电子病历是现代医疗机构开展高效、优质的临床诊疗、科研以及医疗管理工作所必需的重要临床信息资源, 也是居民健康档案的主要信息来源. 海量高质量的电子病历数据蕴含有丰富真实可信的医疗知识和患者的健康信息, 特别是风险因素信息, 如 "既往高血压病史 1 周, 最高可达 180/100 mmHg"、"糖尿病史 10 年" 等, 如能对电子病历中的这些信息进行自动识别抽取, 继而依靠计算机系统长期地对其进行管理控制, 这将对自动化研究心血管疾病的风险因素以及预防心血管疾病具有重大意义.

在自然语言处理领域中, 自动信息抽取主要基于统计机器学习方法, 而由于电子病历文本的半结构化特点和鲜明的子语言特点[3], 使得开放域的标注语料无法适应于电子病历文本的信息抽取. 因此, 一份专门的针对中文电子病历的心血管疾病风险因素标注语料库就成了研究的基础. Stubbs 和 Uzuner[4] 也指出对于特定的医疗信息抽取任务需要构建一个新的标注语料. 语料标注研究作为自然语言处理中重要的组成部分, 一直以来都备受关注: 从持续长达八年之久的 Penn Treebank 词性和句法标注项目[5], 到生物医疗领域涵盖十万级标注的 GENIA 语料库[6], 再到每届 i2b2 (Informatics for integrating biology & the bedside) 特定任务标注语料库.

本文研究的重点是从中文电子病历中挖掘和心血管疾病有关的风险因素, 构建标注体系、标注规范和标注语料库. 我们提出了适用于中文电子病历的心血管疾病风险因素标注体系, 并制定了标注规范, 构建了首份心血管疾病风险因素中文标注语料库.

1　相关研究

1.1　医疗健康信息抽取及语料标注

在医疗信息抽取领域, i2b2 组织的公开评测引起了广泛关注. 在 2006 年组织的对患者吸烟状态的识别任务[7] 中, 患者的吸烟状态被分为五类. 2008 年该评测对电子病历中的肥胖及其并发症进行了抽取[8], 同时在标注时还引入了推断机制, 对检查值如血压值、血糖值等能表征疾病存在的描述同样进行标注. 2009 年关注从电子病历中抽取药物相关的信息[9], 同时药物也是我们对心血管疾病风险因素标注的指针之一. 在 2010 年, i2b2 评测组织者号召参赛者从电子病历中抽取医疗概念、医疗问题的修饰以及医疗问题、检查、治疗之间关系, 同时对检查和药物能够表征医疗问题这一事实也进行了考虑[10]. 在 2012 年的评测任务中, 电子病历中的医疗事件、时间信息以及事件和时间的关系被作为抽取对象[11]. 而 2014 年的任务中则对糖尿病患者电子病历中的心脏病风险因素进行了抽取[4]. 除了 i2b2 之外, 另一些研究者也对健康记录中的医疗术语和实体[12−14]、时间信息[15−16]、医疗事件之间的关系[17−18] 进行了讨论. 同时在医疗信息抽取领域还有一些公开好用的知识库和标注语料库, 如 ICD-9-CM[19]、SNOMED[20]、UMLS[21]、GENIA[6]、Penn Treebank[5].

对于中文医疗信息抽取的研究, 近年来也颇受领域研究

收稿日期 2017-04-17　录用日期 2017-10-29
Manuscript received April 17, 2017; accepted October 29, 2017
国家自然科学基金 (71531007) 资助
Supported by National Natural Science Foundation of China (71531007)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 哈尔滨工业大学计算机科学与技术学院语言技术研究中心网络智能研究室 哈尔滨 150001　2. 哈尔滨医科大学附属第二医院 哈尔滨 150081　3. 哈尔滨理工大学软件学院 哈尔滨 150080
1. Web Intelligence Laboratory, Language Technology Research Center, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001　2. The 2nd Affiliated Hospital of Harbin Medical University, Harbin 150081　3. School of Software, Harbin University of Science and Technology, Harbin 150080

者们的关注. 杨锦锋等[22] 对中文电子病历中的命名实体和实体关系构建了标注语料库, 值得借鉴的是他们采用了预标注训练标注者更新规范的标注模式, 取得了较高的标注一致性. Lei[23] 对中文电子病历中出现的医疗问题、检查、药物、治疗过程进行了抽取研究. Xu 等[24] 则考虑中文电子病历实体识别和分词联合模型. Wang 等[25] 为了从肝癌患者手术记录中抽取与肿瘤相关的信息, 构建了一个含有 961 个和肿瘤相关的实体语料库. Wu 等[26] 用深度学习方法对中文电子病历中的命名实体进行识别, 所用标注语料库为 Lei 等[27] 在 400 份病历上构建的实体标注语料库. Wang 等[28] 对传统中医医疗记录中症状名的标准化进行了探索, 对中医医疗文本中的同个症状的不同症状名进行识别.

1.2 现有的心血管疾病风险因素标注规范和标注语料库

2014 年, i2b2 评测组织者[4, 29] 对糖尿病人的心脏病风险因素进行了考虑. 在制定的标注规范中, 五种风险因素: 高血压、高血脂、吸烟、肥胖和家族史, 两种疾病: 糖尿病和心脏病以及它们的标注原则得到了详细的解释, 首次提出对风险因素和疾病进行指针 (Indicator) 的标注, 指针是一种特征用以表征疾病或风险因素的存在. 在对疾病和风险因素时间属性的标注过程中, 依据的是医疗问题发生的时间和 DCT (Document creation time) 之间的先后关系. 基于构建的标注规范, 评测组织方对 1 304 份具有时间持续性的医疗记录中的疾病和风险因素进行了标注. 最终在具有医学背景的标注者标注下, 构建了首份糖尿病人心脏病风险因素英文标注语料库.

2 中文电子病历心血管疾病风险因素标注体系的建立

以 Gasparyan 描述心血管疾病风险因素的专著[2] 为基础, 结合国家心血管病中心的《中国心血管病报告 2015》、世界心脏病联盟关于心血管疾病风险因素的报告[30] 和世界卫生组织指出的心血管疾病风险因素, 在专业临床医生的指导下, 我们提出了适用于中文电子病历内容特点的 12 种心血管疾病风险因素. 图 1 展示了标注体系缩略图. 下面按风险因素及其指针、时间属性、修饰逐步介绍中文电子病历心血管疾病风险因素的标注体系.

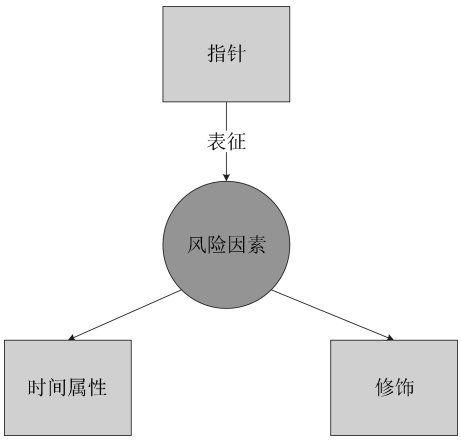

图 1 风险因素标注体系缩略图

Fig. 1 The thumbnail of risk factor annotation scheme

2.1 心血管疾病风险因素及其指针

借鉴 2014 年 i2b2 组织的心脏病风险因素标注任务[4] 经验, 我们制定了中文电子病历心血管疾病风险因素的标注原则, 风险因素的指针是指描述文字是以何种特征映射为相应风险因素. 详见表 1.

我们对出现的相关检查指标值均进行标注, 如 “BP 130/80 mmhg”、“随机血糖: 14.5 mmol/L”, 这样做一是因为未超阈值的检查值在风险因素的发展中有着重要的作用; 二是为了标注方便, 标注者无需对检查值是否超过阈值进行反复判断, 可以节省标注时间, 提高标注效率. 风险因素描述里的能表示程度的信息如持续时间 “高血压病 1 年” 里的 “1 年” 反映风险因素的持续程度、频率 “每天 20 支” 里的 “每天” 表示吸烟频率等都需要一起标注.

2.2 风险因素的时间属性

时间属性表达的是风险因素存在的时间信息. Sun 等在文献 [11] 中指出 “医疗事件的时间信息在病人的诊断、治疗和愈后中起到重要作用”. 时间属性可以分成如下四类: 住院之前、住院期间、出院之后、一直持续. 由于年龄和性别的特殊性, 我们不对这两类风险因素进行时间属性的标注. 对风险因素时间属性的标注需要结合具体的上下文语境和出现的位置.

2.2.1 住院之前

风险因素的发生时间在住院之前, 如以前量过的血压, 以前测过的血糖等. 参见如下例子: 有高血压病史, 最高达 180/100 mmHg (描述中 “180/100 mmHg” 是在高血压病史之后出现, 认为这里的 “最高达” 发生于过去, 对应的风险因素高血压发生于住院时间以前, 标注高血压的时间属性为住院之前).

2.2.2 住院期间

风险因素的发生时间在住院期间, 如住院期间的检查、治疗、症状、确诊的疾病等. 结合我们的病历特点, 对在出院小结中的部分: 门诊收治诊断、临床初步诊断、临床确定诊断、入院时情况中的查体、治疗经过、出院时情况出现的风险因素以及首次病程记录中的部分: 病例特点中的查体、临床初步诊断、诊断依据中的查体、诊疗计划出现的风险因素均标注时间属性为住院期间. 参见如下例子: 查体: Bp 130/80 mmHg (对应的高血压标注时间属性为住院期间).

2.2.3 出院之后

风险因素的发生时间在出院之后, 如医生开出院之后需要服用的药物、对症治疗等. 对在出院小结中的出院医嘱部分出现的风险因素时间属性标注为出院之后. 参见如下例子: 出院医嘱: 控制血压, 监测血糖 (控制血压对应于高血压, 对其时间属性标注为出院之后).

2.2.4 一直持续

风险因素在短时间内不会改变, 持续于住院之前、住院期间、出院之后. 参见如下例子: 身材肥胖, 发育正常 (这里的肥胖对应于超重或肥胖, 一般认为患者的肥胖不会在一次住院期间而改变, 因此标注时间属性为一直持续).

在对风险因素时间属性的标注过程中, 需要注意的是对高血压、糖尿病、血脂异常、慢性肾病、动脉粥样硬化这些慢性疾病的标注, 在用以上规则都无法判断的情况下均标为一直持续, 因为这些疾病病程周期长、治疗过程一般只是起到控制和预防并发症的作用, 较难彻底治愈如初, 会长期的存在于患者身上.

表 1　中文电子病历心血管疾病风险因素的标注原则

Table 1　Annotation guidelines for CVDs risk factors in CEMR

类别	风险因素	指针	标注原则
疾病类	超重或肥胖	病历提到	提到体重超重或者肥胖的描述, 如: 身材**肥胖**
		腰围值	提到患者的腰围或者腹围值
	高血压	病历提到	提到高血压或高血压病史, 如: 既往**高血压病 1 年**(这里带有持续时间我们将其一同标注)
		血压高	提到患者的血压值或任何反映患者血压高的表述, 如: 查体: ··· **BP 130/80 mmHg**···
		调节血压	提到患者需要调压或已有调压效果不理想的描述, 如: **血压控制不理想**
		药物	明确目的是为了调压的药物, 如: 平素口服**珍菊**降压
	糖尿病	病历提到	提到糖尿病或糖尿病病史, 如: 无**糖尿病病史**
		血糖高	提到血糖高、血糖的相关检查指标值或者其他可以表明患者血糖高的描述, 如: **随机血糖: 14.5 mmol/L**
		调节血糖	提到患者需要调节血糖或已有调节效果不理想的描述, 如: 长期以来血压、**血糖控制不佳**
		药物	明确目的是为了调节血糖的药物、饮食, 如: 口服**降糖药**控制尚可
	血脂异常	病历提到	提到血脂异常、高血脂或高血脂史, 如: **高血脂 10 余年**
		血脂高	提到患者血脂的相关检查指标值或任何可以表明患者血脂高的描述, 如: **总胆固醇 (GPO 酶法): 5.39 mmol/L**
		调节血脂	提到患者需要调脂或已有调脂效果不理想的描述, 如: 诊疗计划: **控制血脂**
		药物	明确目的是为了调脂的药物, 如: 调节血脂, 稳定冠脉粥样斑块: **立普妥** 20 mg Qn po
	慢性肾病	病历提到	提到慢性肾病的描述, 如: 病历特点: **肾炎病史 20 余年**
	动脉粥样硬化	病历提到	提到动脉粥样硬化、粥样斑块或冠脉狭窄的描述, 如: 临床确定诊断: **冠状动脉粥样硬化**
	阻塞性睡眠呼吸暂停综合征	病历提到	提到阻塞性睡眠呼吸暂停综合征的描述, 如: 临床确定诊断: 肾囊肿**阻塞性睡眠呼吸暂停综合征**
生活方式类	吸烟	病历提到	提到患者吸烟或吸烟史的描述, 如: **吸烟 40 余年**
		戒烟	提到患者戒烟或未戒烟的描述, 如: **戒烟**1 年 (这里的 1 年表示戒烟距现在的时常, 不代表吸烟的时间长短, 因此不能反映吸烟的严重程度)
		吸烟量	提到患者吸烟量的描述, 如: ···**每天 20 支**
	过度饮酒	病历提到	提到患者过度饮酒或饮酒严重程度的描述, 如: **嗜酒 40 余年**
		饮酒量	提到患者饮酒量的描述如: 饮酒史 20 余年, **1 斤/日**
	心血管疾病家族史	病历提到	提到患者有心血管疾病家族史或一级亲属 (父母、兄弟姐妹、子女) 有心血管疾病史, 如: 病例特点: ···**母亲患有冠心病**···
不可改变类	年龄	病历提到	提到患者的年龄, 如: 年龄: **55 岁**
		年龄层	提到患者所处的年龄层, 如: **青年**男患
	性别	病历提到	提到患者的性别, 如: 中年**男**患

2.3　风险因素的修饰

风险因素标注体系中, 我们对风险因素进行修饰的判断, 将风险因素按是否对患者本人进行考虑以及对患者本人进行考虑时发生的确定程度进行如下分类: 肯定的、否定的、可能的、非患者本人的. 由于性别和年龄的特殊性, 不对其进行修饰的标注.

2.3.1　肯定的

风险因素确定发生过或正在发生于患者身上, 则标注为肯定的. 比如下面的例子: 既往高血压病史 20 年 (高血压病史 20 年确定在患者身上发生过, 修饰标注为肯定的).

2.3.2　否定的

风险因素在患者身上进行了考虑, 但确定不在患者身上发生. 一般表述前有否定词如 “无”、“否认”、“未” 等来否定患者风险因素的存在. 比如下面的例子: 既往否认高血压、糖尿病史 (这里的 “高血压” 和 “糖尿病史” 对应风险因素高血压和糖尿病, 修饰标注为否定的).

2.3.3　可能的

风险因素在患者身上进行了考虑, 但不确定在患者身上是否发生, 需要进一步的证据才能确定, 如风险因素出现在出院小结中的门诊收治诊断、临床初步诊断部分, 或者出现在病程记录的临床初步诊断等诊断信息待确定的部分. 比如下面的例子: 临床初步诊断: ··· 糖尿病 (这里的 “糖尿病” 为初步诊断结果, 仍然需要进一步确诊, 故对风险因素糖尿病标注修饰为可能的).

2.3.4　非患者本人的

风险因素不在患者身上进行考虑, 如发生在患者的亲属或者别人身上. 比如下面的例子: 父亲有高血压、冠心病病史 (这里的 “高血压” 对应风险因素高血压, 但不是在患者身上发生的, 故标注为非患者本人的).

2.4　与已有标注体系的对比

本文的标注体系借鉴了 2014 i2b2/UTHealth 心脏病风险因素的标注体系, 但在此基础之上又有如下的改进和补充: 1) 结合诸多权威心血管疾病风险因素研究, 基于中文电子病历的特点, 在专业医生的指导和参与下, 我们提出了适合中文电子病历的 12 种心血管疾病风险因素标注原则, 相比较于 2014 i2b2/UTHealth 中标注心血管疾病的子集心脏病的 5 种风险因素, 种类范围更加完备; 2) 对于风险因素的相关检查值, 不论是否超过阈值, 我们均对其进行标注. 2014 i2b2/UTHealth 中只对超过阈值的检查值进行风险因素标注, 这就缺少对未超阈值信息的抽取并增加了标注难度; 3) 对于一个长期风险因素监控系统来说, 风险因素的时间信息至关重要. 在 2014 i2b2/UTHealth 中考虑的是风险因素和 DCT 之间的关系, 而中文电子病历中的入院和出院时间是被详细记录下来的, 根据住院时间来确定风险因素的发生时间是最适合中文电子病历心血管疾病风险因素标注的; 4) 在 2014 i2b2/UTHealth 中, 心脏病风险因素的标注并未考虑修饰信息, 也就无法确定风险因素发生的确定程度, 这在我们的标注体系中得到了补充.

3　中文电子病历心血管疾病风险因素标注规范的制定和语料库构建

依据上面建立的中文电子病历心血管疾病风险因素标注体系, 我们构建了标注规范初稿, 开发了标注工具, 对经过处

理后的数据进行了标注. 我们邀请一名专业医生 (哈尔滨医科大学附属第二医院呼吸内科住院医生、医学博士) 参与规范的制定和完善, 同时聘请两名医学在读硕士生 (哈尔滨医科大学) 参与病历的标注. 整个语料构建的过程可以分为三个阶段, 具体的构建实施流程见图 2, 以下对其中的一些细节进行介绍.

3.1 数据准备

我们从哈尔滨医科大学附属第二医院 2012 年全年的共 140 000 份病历包括心血管内科、心血管外科等在内的 35 个大科室 87 个子科室中筛选出 600 名患者的出院小结和首次病程记录共 1 200 份病历作为使用数据集. 这 600 名患者中有 344 名来自心血管内科, 190 名来自心血管外科, 还有 66 名来自其他科室. 由于医院保存的病历是快照的形式, 为了对数据进行标注, 需要对图片格式的病历进行文本化操作: 1) 用光学字符识别 (OCR) 工具 “Tesseract”[31] 将图片格式的电子病历转化为文本格式; 2) 人工修正 1) 中自动识别时产生的信息错误, 同时移除掉隐私信息, 如病人的姓名、家庭住址、住院号、医生姓名等; 3) 对修正及去隐私后的病历进行 XML 编码, 用 XML 标记为病历每一部分添加子标题.

3.2 标注规范的制定和标注人员的培训

根据标注体系制定规范初稿, 同时应用开发的风险因素标注工具对如上经过预处理的电子病历进行风险因素的标注. 标注包括对风险因素类型、指针、时间属性和修饰进行

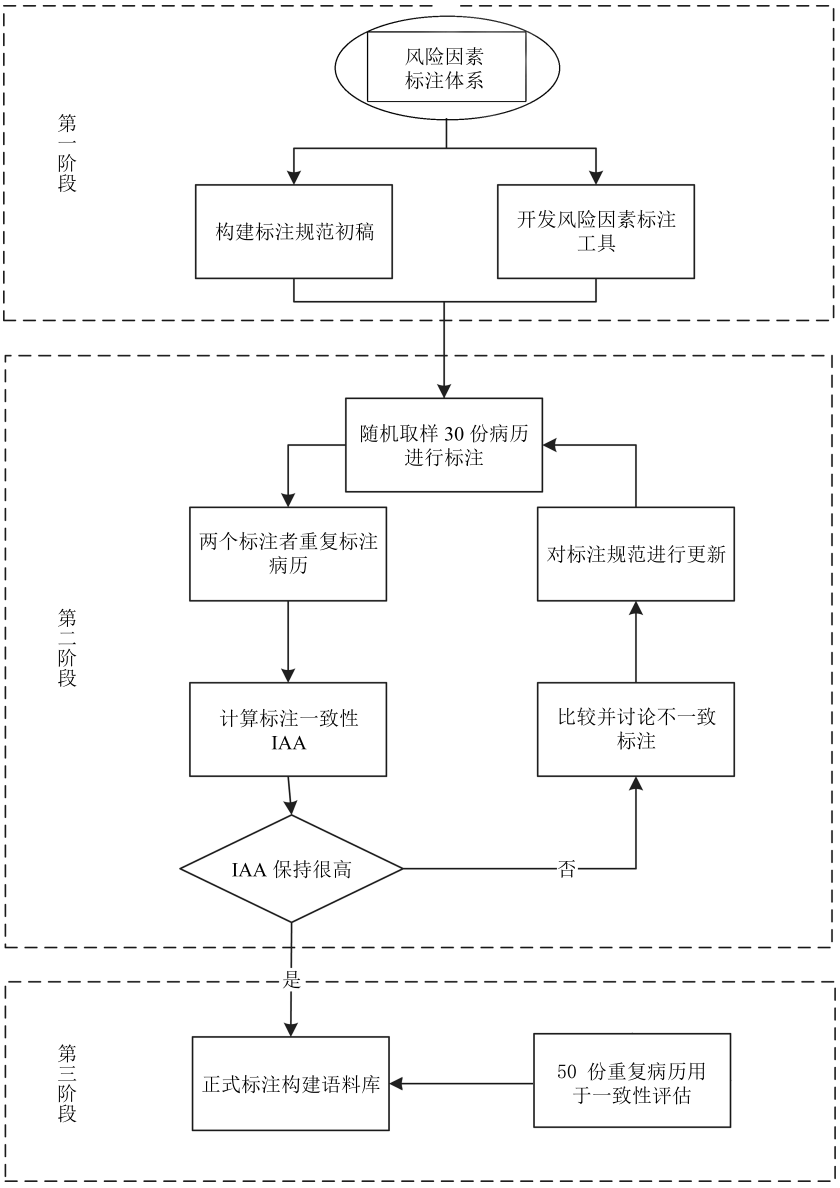

图 2 风险因素语料构建流程图

Fig. 2 Annotation flow chart of risk factor corpus

分类, 同时对于不确定的标注打上不确定标签. 之后, 我们对两位标注者的标注进行一致性 IAA (Inter-annotator agreement) 计算, 并与标注者讨论不一致标注, 对规范中存在的问题进行修改和更新.

如上的培训过程迭代执行, 直到第二步中两位标注者的标注一致性能够持续地达到很高的水平. 最终我们一共进行了 5 轮培训, 每轮的标注数据均不相同, 而在每一轮中两位标注者标注相同的 15 份病历, 一致性评价结果如表 2 所示.

文献 [32] 指出, 当标注一致性达到 0.8 时, 即可认为语料的一致性是可信赖的. 在五轮标注培训之后, 我们认为标注规范和两位标注者已满足正式标注语料库的要求.

表 2　五轮培训的标注一致性

Table 2　The IAA in the training

	第一轮	第二轮	第三轮	第四轮	第五轮
P	0.810	0.977	0.967	0.986	0.988
R	0.815	0.977	0.962	0.986	0.988
F	0.812	0.977	0.964	0.986	0.988

3.3　语料库的构建及结果分析

在正式语料库构建的阶段, 我们采用三种措施来保障标注的质量.

1) 在开发标注工具的时候, 我们添加了一个 “不确定” 标记. 当标注者不确定自己的标注时, 可为该标注打上不确定标签, 用于后续的讨论.

2) 正式标注时, 我们给两位标注者分发的数据中有 50 份病历是重复的, 这部分重复病历可以用来评价标注一致性, 衡量标注语料库的质量.

3) 标注者定期的对标注结果进行提交, 语言学家对标注结果进行随机抽样检查, 保证至少 1/3 的语料被抽查到.

最终, 我们用 1 200 份病历构建了心血管疾病风险因素标注语料库, 总共包含 9 678 个风险因素标注, 标注一致性 F 值为 0.968 表明了该语料库的高质量. 我们对 12 种风险因素的标注结果进行了统计, 具体可见表 3.

表 3　语料库中各风险因素数量统计

Table 3　The statistics of risk factor annotated corpus

类型	风险因素	数量
疾病类	超重或肥胖	18
	高血压	3 729
	糖尿病	1 007
	血脂异常	372
	慢性肾病	26
	动脉粥样硬化	144
	阻塞性睡眠呼吸暂停综合征	1
行为和生活方式类	吸烟	508
	过度饮酒	95
不可改变类	心血管疾病家族史	10
	年龄	1 859
	性别	1 909

4　讨论

在我们的标注体系中, 对心血管疾病的 12 种风险因素进行了考虑, 其中对疾病类的 7 种风险因素如超重或肥胖、高血压、糖尿病等的标注, 可以用来构建抽取模型自动抽取患者的疾病、药物、检查值等信息, 以对这些慢性疾病进行健康管理, 包括检查值的变化情况、药物的有效性监控、疾病的治愈分析等. 在 2014 i2b2/UTHealth 心脏病风险因素标注中, Stubbs 和 Uzuner[4] 即对时间属性的作用做了解释 “为了便于呈现患者心脏病的长期病情变化情况, 而将风险因素和病人的医疗时间点关联起来”. 本文提出的将心血管疾病风险因素和住院时间相关联, 可以直观地看到风险因素随时间变化情况, 如风险因素肥胖在某次住院之前发生了, 而在下次住院时该风险因素是否定的, 第三次住院时肥胖发生在住院之前, 则可以看到该风险因素从出现到消失再到出现的概况. 标注语料库的统计结果表明, 诸如高血压、糖尿病、年龄、性别等是中文电子病历所关注记录的信息, 而对于患者的心血管疾病家族史、慢性肾病、超重或肥胖等信息记录较少, 和前几个风险因素相比较少受到关注. 在对病历的标注过程中, 我们发现中文电子病历中像非健康饮食、缺乏运动等均未出现, 这对于训练风险因素的抽取模型是不利的. 后续如果扩展数据来源到别的数据上如体检报告, 还需要进一步地挖掘这些风险因素.

对数据的初步使用探索表明了风险因素的实用性. 我们对心血管疾病进行了自动诊断实验, 根据患者的病情描述提取文本特征, 通过训练机器学习模型来对文本进行心血管疾病诊断分类, 分为患有和未患有. 数据使用 583 份已标注好医疗实体 (疾病、症状、检查、治疗) 的首次病程记录, 其中心血管疾病患病正例有 235 份. 我们对病历中的实体提取出自述症状、检查结果, 同时使用词典加规则的方法提取患者的心血管疾病风险因素, 对比了机器学习方法 LR (Logistic regression), RF (Random forest), GBDT (Gradient boosting decision tree), XGboost[33]. 实验结果如表 4 所示, 其中的值表示分类的 F 值. 实验表明在加入风险因素信息如高血压、糖尿病、吸烟等之后, 分类效果均有显著提升, 这表明了风险因素信息的实用价值. 我们对风险因素进一步的实用性探索正在进行中.

表 4　心血管疾病诊断实验结果

Table 4　Diagnosis results of CVDs

特征 \ 方法	LR	RF	GBDT	XGboost
自述症状 + 检查结果	0.662	0.672	0.756	0.720
自述症状 + 检查结果 + 风险因素	0.675	0.688	0.798	**0.811**

5　结束语

本文主要总结了中文电子病历心血管疾病风险因素标注问题, 分为以下三部分进行了介绍: 1) 建立适用于中文电子病历的心血管疾病风险因素标注体系; 2) 制定中文电子病历心血管疾病风险因素标注规范; 3) 依据规范, 构建首份中文电子病历心血管疾病风险因素标注语料库. 在标注体系的建立、标注规范的制定完善、语料标注问题的讨论中均有专业医生的指导和参与, 所建立起的标注准则具有较高的科学价值. 在标注的过程中, 我们先对标注者进行了 5 轮培训, 结果显示标注一致性能保持在很高的水平, 最后构建的语料库标注一致性 F 值为 0.968 表明了标注语料库的高质量. 本文提出了如下几点创新: 1) 建立了首个适用于中文电子病历的心血管疾病风险因素标注体系, 包含 12 种风险因素及其相关信息的标注原则; 2) 在对风险因素检查值标注时, 考虑我们的任务需要, 无论数值是否高于判断阈值均对其进行标注; 3) 结合中文电子病历的特点, 引入四类时间属性的标注, 用于观察风险因素和住院时间的关系; 4) 我们对

风险因素的修饰进行了判断, 根据风险因素是否在患者身上进行考虑以及在患者身上进行考虑时的确定程度进行了分类. 所构建的语料库可以用来做风险因素的自动抽取研究以及构建适用于监控中国人健康状况的心血管疾病风险因素自动化管理系统, 预防风险因素的发展, 减少心血管疾病的发生, 提高医疗资源利用率降低医疗成本, 为健康事业做出贡献, 并将对其他类似疾病的预防研究起到积极的参考作用. 同时, 我们的研究是基于国内三甲医院真实的住院病历而开展的, 具有较强的实用性和权威性, 为中文电子病历健康大数据的自动挖掘和利用带来了生机. 研究中制定的规范、开发的标注工具、标注的数据样例以及医院的伦理证明可见 https://github.com/WILAB-HIT/RiskFactor.

致谢

感谢哈尔滨医科大学附属第二医院病案室对本研究的支持. 感谢标注者认真仔细的工作.

References

1 World Health Organization. Cardiovascular diseases (CVDs) [Online], available: http://www.who.int/mediacentre/factsheets/fs317/en/, November 3, 2017.

2 Gasparyan A Y. *Cardiovascular Risk Factor*. Rijeka, Croatia: InTech, 2012. 1−102

3 Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 2002, **35**(4): 222−235

4 Stubbs A, Uzuner Ö. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*, 2015, **58**(S): S78−S91

5 Marcus M P, Marcinkiewicz M A, Santorini B. Building a large annotated corpus of English: the Penn Treebank. *Computational linguistics*, 1993, **19**(2): 313−330

6 Kim J D, Ohta T, Tateisi Y, Tsujii J. GENIA corpus-semantically annotated corpus for bio-textmining. *Bioinformatics*, 2003, **19**(S1): i180−i182

7 Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 2008, **15**(1): 14−24

8 Uzuner Ö. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 2009, **16**(4): 561−570

9 Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 2010, **17**(5): 514−518

10 Uzuner Ö, South B R, Shen S Y, DuVall S L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 2011, **18**(5): 552−556

11 Sun W Y, Rumshisky A, Uzuner Ö. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 2013, **20**(5): 806−813

12 Pradhan S, Elhadad N, South B R, Martinez D, Christensen L, Vogel A, Suominen H, Chapman W W, Savova G. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 2015, **22**(1): 143−154

13 Meystre S M, Kim Y, Gobbel G T, Matheny M E, Redd A, Bray B E, Garvin J H. Congestive heart failure information extraction framework for automated treatment performance measures assessment. *Journal of the American Medical Informatics Association*, 2017, **24**(e1): e40−e46

14 Ford E, Carroll J A, Smith H E, Scott D, Cassell J A. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 2016, **23**(5): 1007−1015

15 Styler IV W F, Bethard S, Finan S, Palmer M, Pradhan S, de Groen P C, Erickson B, Miller T, Lin C, Savova G, Pustejovsky J. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2014, **2**: 143−154

16 Bethard S, Savova G, Chen W T, Derczynski L, Pustejovsky J, Verhagen M. Semeval-2016 task 12: clinical tempeval. In: Proceedings of the 2016 SemEval. San Diego, USA: SemEval, 2016. 1052−1062

17 Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Setzer A. Semantic annotation of clinical text: the CLEF corpus. In: Proceedings of the 2008 LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining. Marrakech, Morocco: LREC, 2008. 19−26

18 Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 2011, **18**(5): 594−600

19 Quan H D, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J C, Saunders L D, Beck CA, Feasby T E, Ghali W A. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, 2005, **43**(11): 1130−1139

20 Stearns M Q, Price C, Spackman K A, Wang A Y. SNOMED clinical terms: overview of the development process and project status. In: Proceedings of the 2001 AMIA Symposium. Washington DC, USA: AMIA, 2001. 662−666

21 Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 2004, **32**(S1): D267−D270

22 Yang Jin-Feng, Yu Qiu-Bin, Guan Yi, Jiang Zhi-Peng. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Automatica Sinica*, 2014, **40**(8): 1537−1562
(杨锦锋, 于秋滨, 关毅, 蒋志鹏. 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 2014, **40**(8): 1537−1562)

23 Lei J B. Named Entity Recognition in Chinese Clinical Text [Ph. D. dissertation], The University of Texas, USA, 2014.

24 Xu Y, Wang Y, Liu T, Liu J, Fan Y, Qian Y. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. *Journal of the American Medical Informatics Association*, 2014, **21**(e1): e84−e92

25 Wang H, Zhang W D, Zeng Q, Li Z F, Feng K Y, Liu L. Extracting important information from Chinese Operation Notes with natural language processing methods. *Journal of Biomedical Informatics*, 2014, **48**: 130−136

26 Wu Y H, Jiang M, Lei J B, Xu H. Named entity recognition in Chinese clinical text using deep neural network. *Studies in Health Technology & Informatics*, 2015, **216**: 624−628

27 Lei J B, Tang B Z, Lu X Q, Gao K H, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association*, 2014, **21**(5): 808−814

28 Wang Y Q, Yu Z H, Chen L, Chen Y H, Liu Y G, Hu X G. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study. *Journal of Biomedical Informatics*, 2014, **47**: 91−104

29 Stubbs A, Kotfila C, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of Biomedical Informatics*, 2015, **58**(S): S67−S77

30 World Heart Federation. Cardiovascular disease risk factors [Online], available: https://www.world-heart-federation.org/resources/risk-factors/, March 28, 2017.

31 Tesseract[Online], available: https://github.com/tesseract-ocr, November 3, 2017.

32 Artstein R, Poesio M. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 2008, **34**(4): 555−596

33 Chen T Q, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: ACM, 2016. 785−794

苏 嘉 哈尔滨工业大学博士研究生. 主要研究方向为信息抽取和自然语言处理. E-mail: sjd163mail@163.com
(**SU Jia** Ph. D. candidate at Harbin Institute of Technology. His research interest covers information extraction and NLP.)

何 彬 哈尔滨工业大学博士研究生. 主要研究方向为命名实体识别, 实体关系抽取. E-mail: hebin_hit@foxmail.com
(**HE Bin** Ph. D. candidate at Harbin Institute of Technology. His research interest covers named entity recognition, entity relation extraction.)

吴 昊 哈尔滨医科大学附属第二医院硕士研究生. 主要研究方向为血管瘤和 circRNA 在纤维化中的作用机制.
E-mail: rosiewuyanxi@gmail.com
(**WU Hao** Master student at the Second Affiliated Hospital of Harbin Medical University. Her research interest covers hemangioma, the modulation mechanism of circular RNA on expressions of fibrosis-associated process.)

杨锦锋 哈尔滨理工大学讲师, 博士. 主要研究方向为健康信息学, 自然语言处理. E-mail: fondofbeyond@163.com
(**YANG Jin-Feng** Lecturer and Ph. D. at Harbin University of Science and Technology. His research interest covers health informatics and NLP.)

关 毅 哈尔滨工业大学教授, 博士. 主要研究方向为健康信息学, 自然语言处理. 本文通信作者. E-mail: guanyi@hit.edu.cn
(**GUAN Yi** Professor and Ph. D. at Harbin Institute of Technology. His research interest covers health informatics and NLP. Corresponding author of this paper.)

姜京池 哈尔滨工业大学博士研究生. 主要研究方向为医疗知识网络, 知识图谱. E-mail: jiangjingchi0118@163.com
(**JIANG Jing-Chi** Ph. D. candidate at Harbin Institute of Technology. His research interest covers medical knowledge network, knowledge graph.)

王焕政 哈尔滨工业大学硕士研究生. 主要研究方向为知识挖掘, 自然语言处理. E-mail: whz123_hit@163.com
(**WANG Huan-Zheng** Master student at Harbin Institute of Technology. His research interest covers knowledge mining, and natural language processing.)

于秋滨 哈尔滨医科大学附属第二医院副主任医师. 主要研究方向为电子病案的数据挖掘. E-mail: yuqiubin6695@163.com
(**YU Qiu-Bin** Deputy chief physician at the Second Affiliated Hospital of Harbin Medical University. Her research interest covers data mining on electronic medical records.)