Check for updates

Received: 2 June 2021 | Accepted: 31 October 2021

DOI: 10.1002/int.22748

RESEARCH ARTICLE

WILEY

A precision-preferred comprehensive information extraction system for clinical articles in traditional Chinese Medicine

Ye Xia[1] | **Jianxiong Cai**[2,3] | **Yizhen Li**[1] | **Zhili Dou**[1] | **Yunan Zhang**[1] | **Lin Wu**[1] | **Zhe Huang**[1] | **Shujing Xu**[1] | **Jiayi Sun**[1] | **Yixing Liu**[4] | **Darong Wu**[2,3,5] | **Dongran Han**[1]

[1]School of Life and Science, Beijing University of Chinese Medicine, Beijing, China

[2]State Key Laboratory of Dampness Syndrome of Chinese Medicine, Guangzhou, China

[3]The Second Affiliated Hospital of Guangzhou, University of Chinese Medicine (Guangdong Provincial Hospital of Chinese Medicine), Guangzhou, China

[4]School of Management, Beijing University of Chinese Medicine, Beijing, China

[5]Guangdong Provincial Key Laboratory of Clinical Research on Traditional Chinese Medicine Syndrome, Guangzhou, China

Correspondence

Dongran Han, School of Life and Science, Beijing University of Chinese Medicine, Room 542, Scientific Research Bldg, Yangguang South Street, Fangshan District, 102400 Beijing, China.
Email: handongr@gmail.com

Darong Wu, State Key Laboratory of Dampness Syndrome of Chinese Medicine, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, #111 Dade Rd, Yuexiu District, 510120 Guangzhou, China.
Email: darongwu@gzucm.edu.cn

Yixing Liu, School of Management, Beijing University of Chinese Medicine, Room 542, Scientific Research Bldg, Yangguang South Street, Fangshan District, 102400 Beijing, China.
Email: yixingliu1987@gmail.com

Abstract

This study established a precision-preferred system specially designed for the data extraction of traditional Chinese medicine (TCM) articles, providing foundational data for subsequent clinical article analysis and synthesis of TCM clinical evidence. Information extraction is commonly used in many fields to identify relevant concepts and the relationship between pairs of concepts from the vast information sources. Previous studies that performed information extraction primarily focused on scattering targeted fields to achieve a balance between precision and recall. Therefore, this study aims to create a comprehensive information extraction system for TCM articles. This system will extract all relevant information from research articles on a broad research field, including the 11 diseases that

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Funding information
National Key R&D Program of China, Grant/Award Numbers: 2019YFC1709801, 2019YFC1709800; State Key Laboratory of Dampness Syndrome of Chinese Medicine, Grant/Award Numbers: SZ2020ZZ09, NO.SZ2021ZZ30

can be efficiently treated with TCM, with high precision and efficient measurement to address bias in every study. It covers the most essential information related to patients, interventions, comparisons, outcomes, and study design (PICOS) principles in TCM clinical trials. This system covers 34 target fields on 14 topics. Impediments such as the various typesetting of TCM clinical articles were managed by a hybrid of machine vision and optical character recognition. Thus, TCM researchers can be spared of laborious, unscalable, and inefficient manual extraction processes. Our system could also enhance TCM researcher awareness of frequently missing information or TCM clinical trial design methods that could introduce bias, by analyzing the overall information integrity of TCM clinical articles, which is beneficial for future research designs.

KEYWORDS
information extraction system, precision-preferred, TCM clinical articles, typesetting recognition

1 | INTRODUCTION

Almost two million academic papers are published yearly, worldwide. For biomedical research alone, one million articles enter PubMed yearly.[1] Such a massive influx of new information makes it difficult for scientists from different fields to "catch up" with developments in their fields. Evidence-based medicine (EBM) utilizes the insights gained through the integration of clinician experience, patient values, and scientific information. To obtain the best scientific information for a given research theme, many studies have been conducted by different researchers in different locations with different patients to avoid bias, resulting in a large number of articles. The United States Food and Drug Administration for pharmaceutical companies requires proof of efficacy of new drugs, which also contributes to the need for many large-scale randomized trials.[2]

In an attempt to handle the massive influx of new scientific information, information extraction technologies have been applied in various fields. This process primarily applies machine learning algorithms with natural language processing techniques to extract the scientific content of interest from articles and aid researchers searching for specific topics. For example, Iris.ai, an article exploration tool, allows researchers to use 300–500 words or the URL of an existing paper to obtain a map of thousands of matching scientific works. Similar tools, such as Semantic Scholar and Microsoft Academic, also provide scientists with a different assessment of scientific articles in a specific niche compared with conventional search engines,

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

such as PubMed and Google Scholar.[3] In addition to these widely used products, researchers also use information extraction technologies to search for information on broad categories of concepts and then identify subsequent relationships between pairs of these concepts.[4] For example, capturing of complex material on science concepts such as the thermoelectricity of lithium-ion cathode materials[5] or collecting evidence and establishing knowledge to support clinicians and clinical research against the coronavirus disease (COVID-19) pandemic.[6]

Commonly applied information extraction technology primarily reduces information overload using two methods: (1) concept extraction that identifies concepts of interest or named entities (e.g., names or addresses),[7] and (2) relationship extraction, which identifies relationships between pairs of concepts of interest.[8] This process is typically used to provide users with recommendations on a series of related articles or possible hypotheses. Given that articles that are falsely extracted can be discarded manually, and that a possible hypothesis is meant to be examined, extraction errors are acceptable in these applications. Most of these information extraction technologies are designed to achieve a balance between precision (a measure of correctness) and recall (a measure of completeness) (see the Supporting Information file), making most established information systems only suitable for narrow research themes.

In the field of clinical information extraction methodology, researchers usually focus on either health records generated from clinical practice or research articles produced by designed clinical experiments. The legislation of the Health Information Technology for Economic and Clinical Health (HITECH) Act[9] facilitated the adoption of electronic health records in many healthcare institutions. The availability of these records has attracted many researchers interested in automated information extraction from clinical texts.[10] Most electronic health record data are in free-text form, a more natural and expressive method, differing from the semistructured style of research articles.[11] Thus, information extraction technologies involving electronic health records typically concentrate on extracting specific clinical concepts[12,13] (e.g., substance usage and disease status), or extracting the diagnosis and treatment information. This is to establish a structured diagnosis and treatment database with key-value pairs,[14] applying extraction methods based on natural language processing and machine learning.

The exploration of information extraction methodology in traditional Chinese medicine (TCM) is still in its initial stages. A study was conducted to systematically evaluate the metadata integrity of randomized controlled trials (RCTs), which provides a reference for improving the structuration of TCM articles, thus accelerating information extraction in TCM articles.[15] Some technologies have been introduced to extract information on TCM. Natural language processing tools were applied to extract TCM and related biomedical information, including genes and biological pathways, from annotated corpus collected from biomedical literatures, to develop a database named TCMGeneDIT, which provides association information between TCM concepts and biological concepts.[16] Artificial intelligence-based technologies are also introduced in diagnostic procedures to help syndrome differentiation based on the electronic medical records of patients.[17] However, studies focusing on the information extraction of RCT articles in TCM are still lacking.

To keep abreast of the vast information in TCM research articles and provide researchers with a public foundational data for subsequent analysis and synthesis of clinical evidence, we summarized the challenges in information extraction of TCM clinical articles and established a precision-preferred comprehensive information extraction system with its architecture depicted in Figure 1.

Specifically, we first discussed the requirements for information extraction from clinical articles, which might not be completely achieved by commonly applied information extraction

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

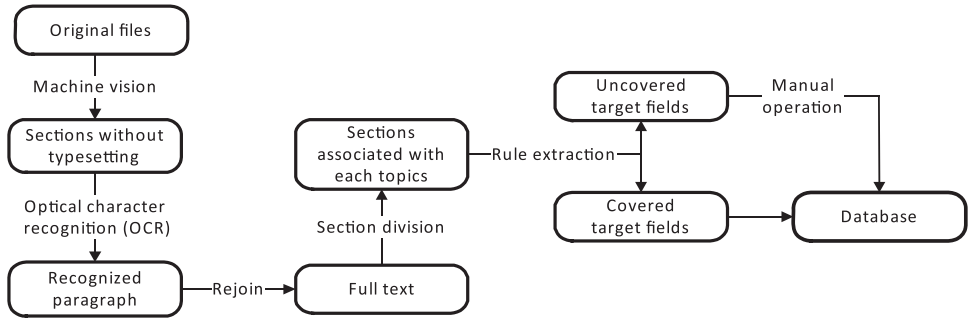


FIGURE 1 The architecture of the information extraction system

technology. In addition to these general requirements, information extraction of TCM clinical articles has additional impediments, such as the need to handle various typesetting and TCM onomatology. Thus, before conducting information extraction for RCT articles in TCM, a hybrid of machine vision and optical character recognition (OCR) was used to transform the original TCM clinical article files into machine-recognizable texts (shown in Figure 2). Then, we present the details of the establishment of the precision-preferred comprehensive information extraction system optimized for Mandarin TCM clinical articles based on a novel extraction rule, which systematically extracts large amounts of information for subsequent analyses, such as systematic reviews and meta-analyses, for a series of narrow research topics.

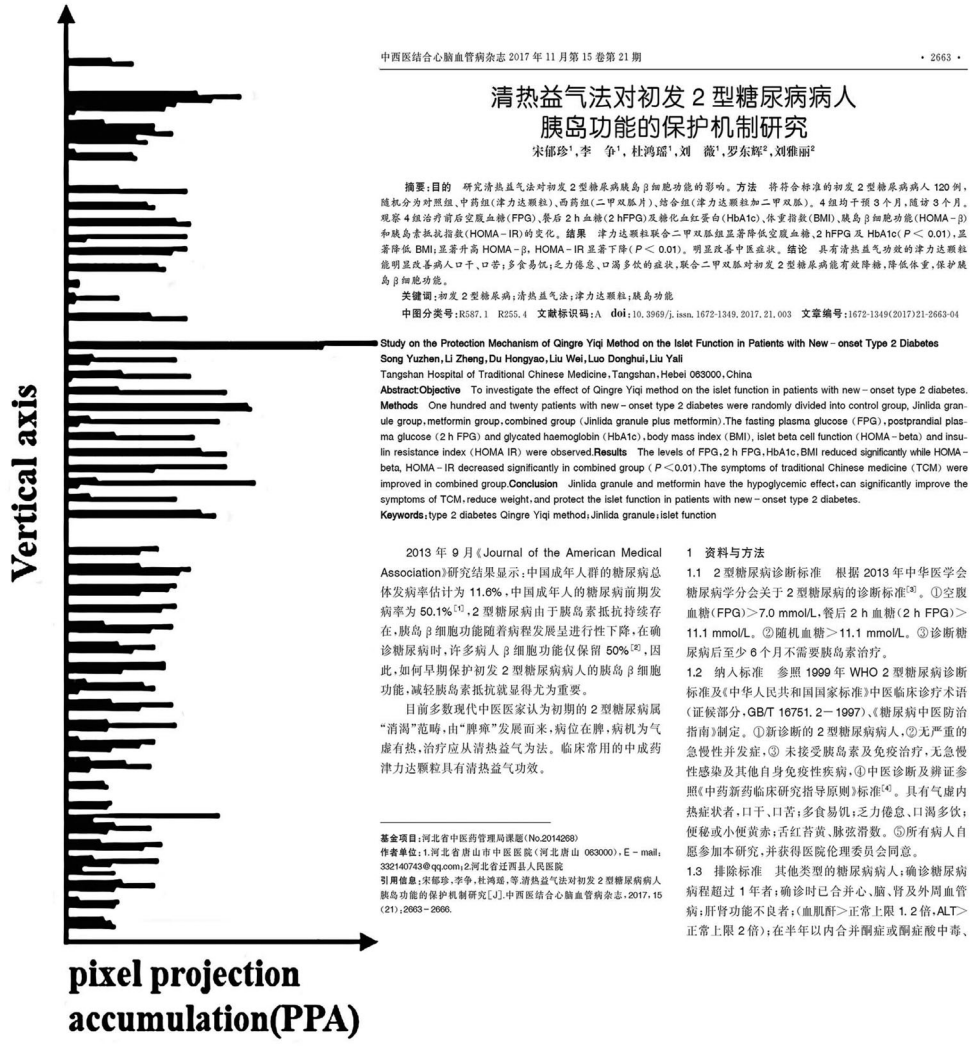
2 | REQUIREMENTS FOR INFORMATION EXTRACTION FROM CLINICAL ARTICLES

Information extraction from clinical articles has special requirements as the data extracted are then used for subsequent synthesis of clinical evidence. This requires a comprehensive assessment of different clinical articles under the same topic from all perspectives to avoid the introduction of bias. Therefore, errors or missing information are considered unacceptable. This requirement makes a comprehensive and precision-preferred information extraction system desirable, instead of current technologies that concentrate on small research topics and pursue a balance between precision and recall. Because most clinical articles are written in a fixed pattern, we found that rule-based information extraction technology may suit the special requirements for information extraction from clinical articles.

Most modern clinical investigations are guided by EBM. According to the Oxford Centre for EBM Levels of Evidence, the highest level of evidence (i.e., level of 1a) requires a systematic review or meta-analysis, such that a consensus can be reached from homogeneous RCTs under the same research topic that may have different or even conflicting results.

Systematic reviews or meta-analyses should assess every topically related RCTs from several perspectives, locating sources of bias, such as selection or performance bias. To determine whether a study should be included, detailed structured information from each article needs to be extracted to perform a systematic assessment. For a given study, the population age and sex distributions are used to determine the heterogeneity of the samples. The applications of random assignment, blinding methods, and allocation concealment are also assessed because of their contribution to the reduction of bias in RCTs. In the traditional manual extraction

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



中西医结合心脑血管病杂志 2017 年 11 月第 15 卷第 21 期 · 2663 ·

清热益气法对初发 2 型糖尿病病人胰岛功能的保护机制研究

宋郁珍[1]，李 挣[1]，杜鸿瑶[1]，刘 薇[1]，罗东辉[2]，刘雅丽[2]

摘要：目的 研究清热益气法对初发 2 型糖尿病胰岛β细胞功能的影响。**方法** 将符合标准的初发 2 型糖尿病病人 120 例，随机分为对照组、中药组(津力达颗粒)、西药组(二甲双胍片)、结合组(津力达颗粒加二甲双胍)。4 组均干预 3 个月，随访 3 个月。观察 4 组治疗前后空腹血糖(FPG)、餐后 2 h 血糖(2 hFPG)及糖化血红蛋白(HbA1c)、体重指数(BMI)、胰岛β细胞功能(HOMA－β)和胰岛素抵抗指数(HOMA－IR)的变化。**结果** 津力达颗粒联合二甲双胍组显著降低空腹血糖、2 hFPG 及 HbA1c($P<0.01$)，显著降低 BMI；显著升高 HOMA－β，HOMA－IR 显著下降($P<0.01$)。明显改善中医症状。**结论** 具有清热益气功效的津力达颗粒能明显改善病人口干、口苦；多食易饥；乏力倦怠、口渴多饮的症状，联合二甲双胍对初发 2 型糖尿病能有效降糖，降低体重，保护胰岛β细胞功能。

关键词：初发 2 型糖尿病；清热益气法；津力达颗粒；胰岛功能

中图分类号：R587.1 R255.4 **文献标识码：**A **doi：**10.3969/j.issn.1672-1349.2017.21.003 **文章编号：**1672-1349(2017)21-2663-04

Study on the Protection Mechanism of Qingre Yiqi Method on the Islet Function in Patients with New－onset Type 2 Diabetes

Song Yuzhen, Li Zheng, Du Hongyao, Liu Wei, Luo Donghui, Liu Yali

Tangshan Hospital of Traditional Chinese Medicine, Tangshan, Hebei 063000, China

Abstract: Objective To investigate the effect of Qingre Yiqi method on the islet function in patients with new－onset type 2 diabetes. **Methods** One hundred and twenty patients with new－onset type 2 diabetes were randomly divided into control group, Jinlida granule group, metformin group, combined group (Jinlida granule plus metformin). The fasting plasma glucose (FPG), postprandial plasma glucose (2 h FPG) and glycated haemoglobin (HbA1c), body mass index (BMI), islet beta cell function (HOMA－beta) and insulin resistance index (HOMA IR) were observed. **Results** The levels of FPG, 2 h FPG, HbA1c, BMI reduced significantly while HOMA－beta, HOMA－IR decreased significantly in combined group ($P<0.01$). The symptoms of traditional Chinese medicine (TCM) were improved in combined group. **Conclusion** Jinlida granule and metformin have the hypoglycemic effect, can significantly improve the symptoms of TCM, reduce weight, and protect the islet function in patients with new－onset type 2 diabetes.

Keywords: type 2 diabetes Qingre Yiqi method; Jinlida granule; islet function

2013 年 9 月《Journal of the American Medical Association》研究结果显示：中国成年人群的糖尿病总体发病率估计为 11.6%，中国成年人的糖尿病前期发病率为 50.1%[1]，2 型糖尿病由于胰岛素抵抗持续存在，胰岛β细胞功能随着病程发展呈进行性下降，在确诊糖尿病时，许多病人β细胞功能仅保留 50%[2]，因此，如何早期保护初发 2 型糖尿病病人的胰岛β细胞功能，减轻胰岛素抵抗就显得尤为重要。

目前多数现代中医医家认为初期的 2 型糖尿病属“消渴”范畴，由“脾瘅”发展而来，病位在脾，病机为气虚有热，治疗应从清热益气为法。临床常用的中成药津力达颗粒具有清热益气功效。

基金项目：河北省中医药管理局课题(No.2014268)

作者单位：1.河北省唐山市中医医院(河北唐山 063000)，E－mail：332140743@qq.com；2.河北省迁西县人民医院

引用信息：宋郁珍，李挣，杜鸿瑶，等.清热益气法对初发 2 型糖尿病病人胰岛功能的保护机制研究[J].中西医结合心脑血管病杂志，2017，15(21)：2663－2666.

1 资料与方法

1.1 2 型糖尿病诊断标准 根据 2013 年中华医学会糖尿病学分会关于 2 型糖尿病的诊断标准[3]。①空腹血糖(FPG)＞7.0 mmol/L，餐后 2 h 血糖(2 h FPG)＞11.1 mmol/L。②随机血糖＞11.1 mmol/L。③诊断糖尿病后至少 6 个月不需要胰岛素治疗。

1.2 纳入标准 参照 1999 年 WHO 2 型糖尿病诊断标准及《中华人民共和国国家标准》中医临床诊疗术语(证候部分，GB/T 16751.2－1997)、《糖尿病中医防治指南》制定。①新诊断的 2 型糖尿病病人，②无严重的急慢性并发症，③ 未接受胰岛素及免疫治疗，无急慢性感染及其他自身免疫性疾病，④中医诊断及辨证参照《中药新药临床研究指导原则》标准[4]。具有气虚内热症状者，口干、口苦；多食易饥；乏力倦怠、口渴多饮；便秘或小便黄赤；舌红苔黄，脉弦滑数。⑤所有病人自愿参加本研究，并获得医院伦理委员会同意。

1.3 排除标准 其他类型的糖尿病病人；确诊糖尿病病程超过 1 年者；确诊时已合并心、脑、肾及外周血管病；肝肾功能不良者；(血肌酐＞正常上限 1.2 倍，ALT＞正常上限 2 倍)；在半年以内合并酮症或酮症酸中毒、

FIGURE 2 A TCM clinical article page sample and its corresponding pixel projection accumulation (PPA) distributions. TCM, traditional Chinese medicine

procedure, hundreds or thousands of related articles on a given research topic are searched and collected. Skilled experts in this topic discuss and reach a consensus on the inclusion and exclusion criteria. After a cursory examination of selected articles using the inclusion and exclusion criteria, detailed information is extracted from each article separately for subsequent statistical analysis. Given that there are usually many research articles on a given topic, the traditional manual extraction method is laborious, unscalable, and inefficient. For example, in a review article on research trends of TCM formula, 26,917 articles were manually extracted, but only 2621 of them were included. The subsequent data extraction process for succeeding analysis is even more time-consuming.[18] In most studies, only the data of interest are extracted, which makes them incomparable to future studies; therefore, the extracted data are mostly discarded after the article is published.

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Automating systematic reviews via machine learning is a burgeoning field, and related studies are performed in a fragmented manner in many aspects of this field. For example, SR Toolbox is a publicly available catalogue of software tools that helps researchers streamline the writing of systematic reviews. Text classification systems for filtering RCTs are mature for practice, and machine learning techniques for identifying RCTs have already been applied in Cochrane.[19] Another machine learning system named Robot Reviewer is used to automatically determine the risk of bias for the domains defined by the Cochrane Risk of Bias tool in clinical trials.[20]

Although the balance of precision and recall serves as a criterion for machine-learning-based information extraction methods, errors may result in misleading conclusions in systematic reviews or meta-analyses of clinical RCTs. For example, false-positive results (i.e., an unacceptable result is incorrectly classified as acceptable) obtained from studies that were not well designed would introduce biases or even errors in the subsequent analysis, which further hampers the effectiveness of clinical evidence.

Unlike the machine-learning-based extraction method, the rule-based extraction method (see detailed introduction in the Supporting Information file) is a declarative approach with a transparent and controllable extraction mechanism, which can provide a near 100% precision rate. Research articles that cannot be extracted by rule-based extraction are explicitly marked for a subsequent manual review, eliminating the possibility of a mixture of correct and false results. Results from rule-based extraction can be stored in a free-text format, which is compatible with research applying machine learning-based methods. Regular expressions have been used to extract specific values such as body weight[21] or blood pressure[22] from electronic health records in previous studies, but the extracted values were not integrated to create a comprehensive system in these studies. In this study, a rule-based extraction method was used to systematically extract information from clinical articles in TCM with high precision to create a comprehensive system that includes all useful information for clinical research and practice.

3 | UNIQUE CHARACTERISTICS OF TCM CLINICAL ARTICLES

In addition to the general requirements of information extraction for clinical articles, some unique aspects need special attention when extracting information from TCM articles.

First, TCM was initially developed in a context with low levels of informatization, with many early publications being manually typeset and published in printed journals. Recently, many of these articles were scanned and transformed into electronic articles. This process is highly dependent on the quality of the initial scan, which leads to a significant variation in the typesetting. In addition, proper electronic indexing is often missing for these transformed articles. When transforming the scanned portable document format (PDF) of these articles into machine-recognizable texts, the commonly applied normal transforming methods cannot be used for TCM articles containing images. Therefore, OCR, a technique used on both characters and images, was applied in this study to transform all articles into machine-recognizable text forms to avoid compatibility issues. Given that most OCR software for Mandarin can only detect simple typesetting (e.g., vertical text), in this study, the articles were first divided into sections with simple typesetting by machine vision before the use of OCR.

Second, most TCM articles are written in Mandarin, which cannot be directly extracted using information extraction techniques designed for the English language. The performance of information extraction techniques in different languages is influenced by factors, such as the

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

linguistic corpus, model design, and several other factors. The linguistic characteristics may differ between the TCM article and common linguistic corpus, and studies concentrating on linguistic characteristic of TCM are lacking, but a widely used multilingual model on a common linguistic corpus indicates that the model has a better performance in English than Mandarin (see Support Information file). In this study, rule-based extraction techniques suitable for Mandarin TCM articles were designed according to the grammar rules of Mandarin.

Third, TCM contains onomatology different from standard Mandarin, requiring special adjustments. For example, TCM describes symptom subtypes using the combination of their spatial status including notions like "biao" (surface) or "li" (inner); temporal status including notions like "yang-ming" (the peak of "yang") and "jue-yin" (when "yin" turns into "yang"); and internal properties including "feng" (wind), "han" (cold), and "shi" (wet), which are not used for expressions in standard Mandarin. Thus, in this study, the rule-based extraction technique was combined with dictionaries created by TCM experts to extract domain-specific languages in TCM.

4 | TRANSFORMATION OF TCM ARTICLES INTO MACHINE-RECOGNIZABLE TEXTS

We downloaded 63,186 Mandarin TCM clinical articles related to 11 target diseases, for which TCM is commonly used as an effective form of treatment. All the original PDF files were transformed into machine-recognizable texts using a hybrid of machine vision and OCR techniques.

The 11 target diseases were chosen from a list of diseases with exceptional TCM treatment effects in the national administration of TCM. These 11 target diseases, including stroke, colorectal cancer, coronary heart disease, heart failure, chronic obstructive pulmonary disease, diabetes, diabetic nephropathy, osteoarthritis, obesity, rheumatoid arthritis, and diarrhea, are prevalent illnesses with extensive interest. We searched and downloaded 63,186 TCM RCT clinical articles on these 11 target diseases from the China National Knowledge Infrastructure (CNKI) for information extraction.

To address the irregular Chinese typesetting of TCM articles, we introduced machine vision to divide articles into sections with simple typesets. PyMuPDF, a GPL-based open-source Python module, was used to transform them into a lossless compression image format (i.e., portable network graphics [PNG]). Then, with a BSD-based open-source Python module, OpenCV, the images underwent grayscale mapping and binarization to form a monochrome image that only contained pure black (grayscale 255) and white (grayscale 0) pixels.

Pure black pixels were projected onto the horizontal and vertical axes to generate a pixel projection accumulation (PPA) distribution. The pixel accumulation peak in the PPA distribution corresponds to the literal lines in the article, while the pixel accumulation corresponds to the intervals between the lines and paragraphs (Figure 2).

In the vertical axis, the articles are composed of literal lines and intervals, both of which contain multiple pixel lines. At the juncture of the literal lines and intervals, the PPA of the pixels changes significantly, indicating the boundary of the literal lines. Pixels, comprising of Chinese characters, are usually distributed more densely in the middle and sparsely on the border. We used the PPA of a given pixel line to subtract the PPA of the following three-pixel lines separately, to retain the maximum outcome. A three-pixel maximum gradient table (TMGT) was generated by traversing through the PPA distribution in a specific direction. By sorting the TMGT in descending order, the maximum gradient change rate can be used as a threshold value to mark the potential boundaries. Traversing the TMGT helps mark all potential boundaries, which may be intervals of

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

lines or sections. In a specific article, the intervals of the sections are wider than the intervals of the literal lines. The width of all continuous blank spaces in the PPA distribution was calculated and sorted, with the median of these widths taken as the threshold of intervals of sections or lines. This allowed for the dissection of the original picture into multiple sections from all the boundaries of intervals (Figure 3).

Each section underwent a similar calculation on the horizontal axis to be dissected into smaller sections. The iterative use of this process continues until no boundaries for intervals of sections could be found. The resulting sections are then composed of a simple typesetting in Mandarin that can be easily recognized (Figure 4). The sections are numbered from top to bottom, left to right, and transformed into text by traditional OCR. Finally, sections are recombined in the order of numeration, leading to the corresponding machine-recognizable text of the article in PDF form.

5 | ESTABLISHING A COMPREHENSIVE INFORMATION EXTRACTION SYSTEM FOR TCM CLINICAL ARTICLES

Regular expression tools were implemented in Python to systematically extract 34 target fields associated with 14 topics according to the fixed descriptive pattern of TCM clinical articles.

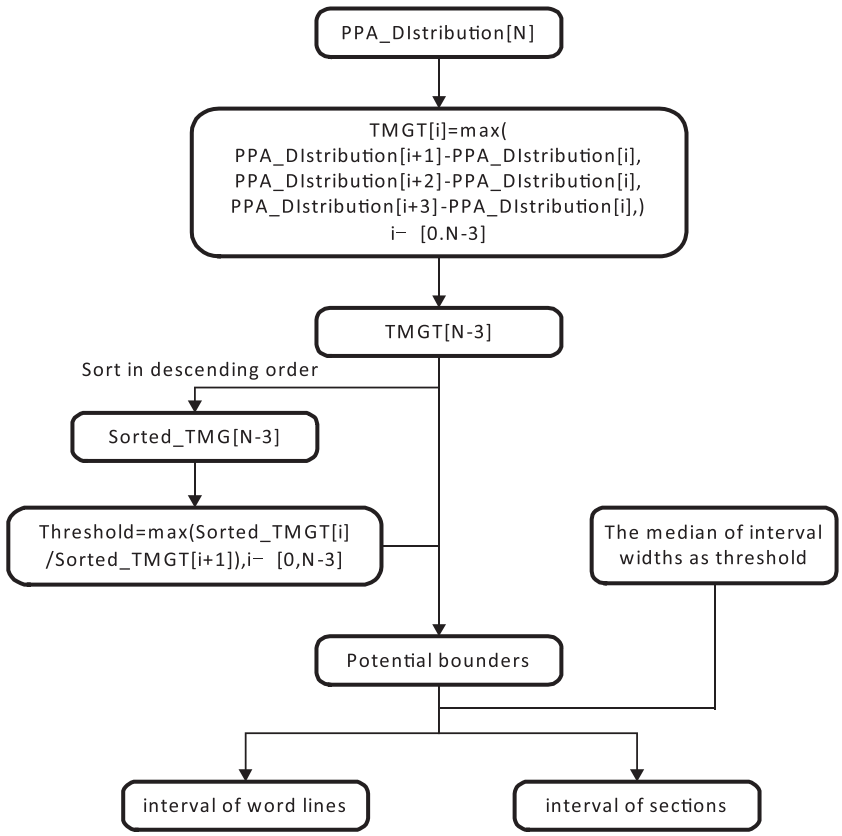


FIGURE 3 The process of recognizing the interval of literal lines and sections in one direction. PPA, pixel projection accumulation; TMGT, three-pixel maximum gradient table

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

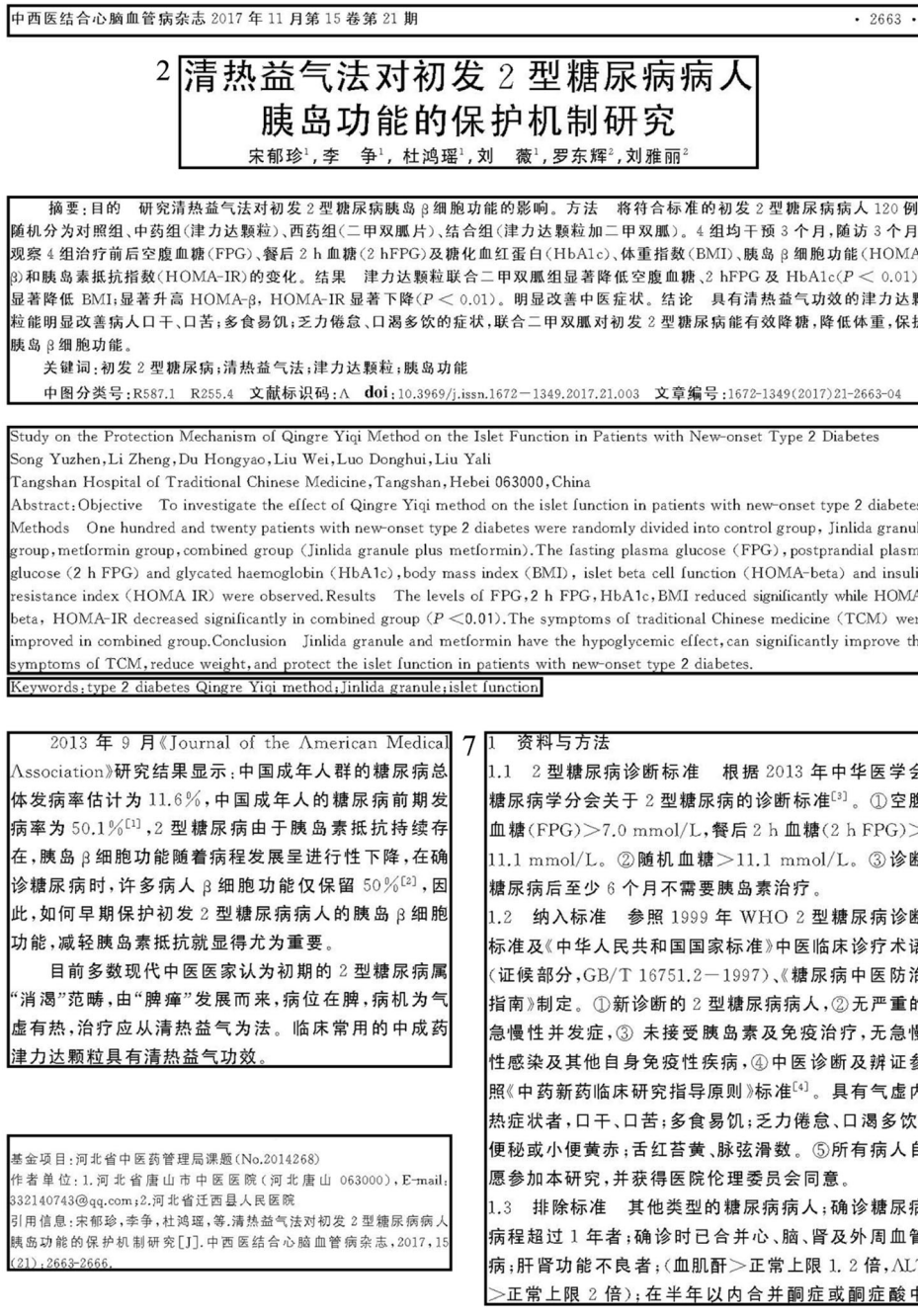


FIGURE 4 Example page is divided into numerical sections

Regular expressions, which are defined as strings of text that allow users to create patterns used for matching, locating, and managing text, are commonly used in standard rule-based information extraction. This tool is already being implemented in Python as a module.

TCM articles are usually written in a fixed semistructured pattern, which is beneficial for the rule-based extraction. Bias-reduction is needed for TCM clinical articles, which requires the

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

assessment of a given study design and specific patterns of its presentation. Most TCM clinical articles are written based on the PICOS model, with the following components: P (patient), I (intervention), C (comparison), O (outcome), and S (study design). This model helps avoid potential biases. The following included six sections: title, abstract, introduction, clinical information, intervention, results, and discussion, each contains certain PICOS-style information. These sections start with fixed patterns and are arranged in this typical order.

In the methods used in this study, the full text underwent sentence segmentation using punctuations, before subsequent extraction. Then, according to the semistructured style of TCM articles, we wrote the initial regular expression patterns to divide the full text into the six sections, with each section associated with the topics that may appear (Figure 5), and each topic containing target fields that describe this topic. If several topics were presented in adjacent locations, a section needed to be further divided into subsections to avoid errors. For example, if the patient cases in both control and experimental groups were described in the clinical information section, this section would be further divided into subsections that separately contained the data of both groups. Then, useful information (e.g., the number of patients) can be extracted from each subsection and labeled separately for the control and experimental groups.

In the rule creation process, the primary goal was to optimize precision (i.e., the proportion of the actual correctly extracted results of all cases classified as correctly extracted results). If only the results from perfectly matched rules are retained, the results typically have a low recall (i.e., the proportion of correctly extracted results of all cases that should have been extracted). For the imperfectly matched results, an iterative approach was used to improve the recall. In this process, a given number of articles was randomly selected as a sample, and the rules were created based on the descriptive patterns in this sample. Then, another sample of articles with the same sample size was selected and extracted using the rules created. In this new sample, if none of the statements in an article was matched by the rules, it could return two different outputs. For example, if the goal was to extract TCM symptom subtypes in an article but the extraction of this article on TCM symptom subtype returned "no match," there may be two

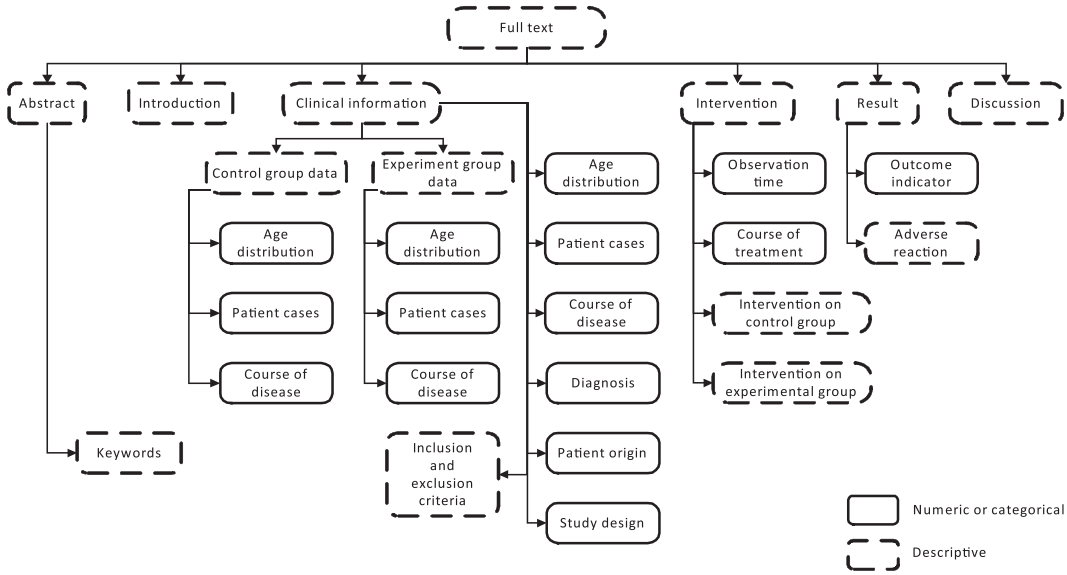


FIGURE 5 The association of each topic with the sections from a typical clinical article

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

possible reasons: (1) the TCM symptom subtypes were not involved; or (2) they were mentioned obliquely using statements that could not be matched by our rules. To determine the source of unmatched results, keywords used to indicate the appearance of TCM symptom subtypes (e.g., "TCM symptom" or "dialectical diagnosis") were searched through all sections that may be associated with TCM symptom subtypes. If the keyword match procedure returned "false," this suggests that the first condition occurred and the target fields were marked as "unmentioned" If the keyword match procedure returned "true," this suggests that the latter condition occurred and the target fields were marked as "manual operation." All target fields marked as "manual operation" in the information extraction result of the new sample were manually checked, which led to the development of new rules suited for their descriptive patterns. If the descriptive patterns were too irregular to be extracted by regular expression, the target field remained listed as "manual operation," which implies that this target field in a specific article is still uncovered and further manual operation might be required. This process iteratively helps achieve acceptable levels of recall (Figure 6).

As mentioned previously, we systematically covered essential information about PICOS for 34 target fields that were associated with 14 topics, which are illustrated in detail below.

Extracted information related to patients included five topic areas: age distribution, sex distribution, sample sizes, course of a disease, and inclusion and exclusion criteria. Patient information for the control and experimental groups was extracted separately. The age distribution and disease course topics included four target fields for extraction: minimum age/course, maximum age/course, average age/course, and age/course variance. Sex distribution topics included two target fields: males and females. The total number of cases included patients who dropped out or did not complete the intervention. Inclusion and exclusion criteria were a description of the criteria used to include or exclude the patients.

Extracted information related to the intervention and control groups included only one topic, the intervention itself. The intervention topic included four target fields: intervention types, intervention subject/intervention acupoint/physiotherapy type, intervention in the control group, and intervention in the experimental group. For ethical reasons, the control group should also receive interventions that have been proven efficacious; therefore, intervention information for both the experimental and control groups was extracted separately. TCM typically includes three types of interventions: medicine, acupuncture, and physiotherapy. Acupuncture can be regarded

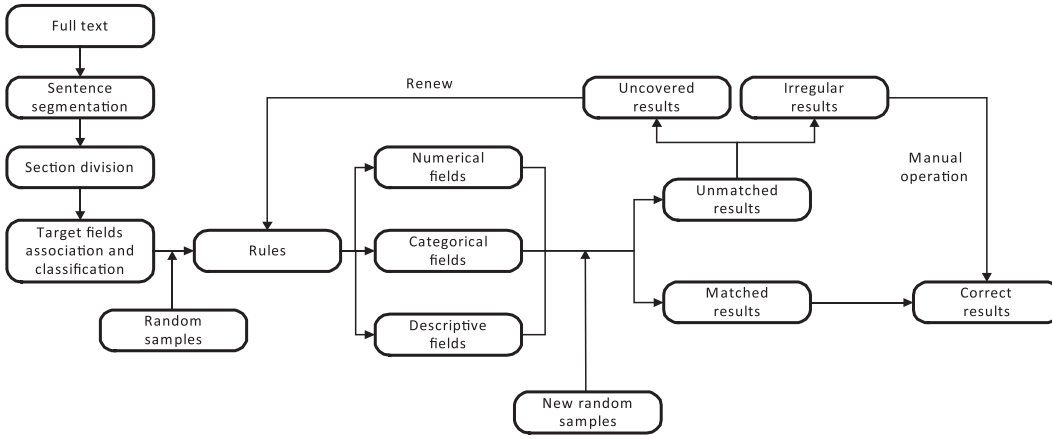


FIGURE 6 The iterative process of rules creation to ensure correct results

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

as a type of physiotherapy, but it plays a more important role than the other physiotherapy types in TCM. Accordingly, many TCM studies use acupuncture as the intervention method, so we listed acupuncture independently as an intervention type. For intervention types, we extracted the specific substance used as medicine, the acupoint in an acupuncture intervention, and the specific physiotherapy type (e.g., "Tai-Chi"). The information on the frequency of intervention was also extracted. Because there may be additional requirements for an intervention, two target fields describing an intervention for the experimental/control groups were extracted; these can be used for further information extraction.

The extracted information on the outcome has two topics: outcome indicator and observation time, with each included as a target field. Outcome indicator is used to measure the efficacy of an intervention, and the observation time indicated how long it took for the outcome indicator to be measured.

The extracted information related to the study design contains two topics: study design and risk of study design. The study design included three target fields: random assignment, blinding, and allocation concealment. Random assignment is a method in which patients are randomly assigned to the experimental or control group. The blinding method refers to the method of preventing a patient and his/her relatives, the doctor, and the data collector from knowing which intervention is being given to this patient. The risk of study design indicates the risk that the random assignment, blinding method, and allocation concealment might lack; this was labeled as high, low, or unclear.

In addition to PICOS-related information, information on three other topics was also extracted. First, keywords, which indicate what the article is focused on according to the authors' opinion were extracted. Second, TCM has its own diagnostic system; thus, diagnosis information, which included five target topics (modern medicine diagnosis, modern medicine diagnostic reference, TCM diagnosis, TCM diagnostic reference [that describes the diagnosis uniquely], and the diagnosis reference of modern medicine or TCM medicine), was extracted. In the TCM diagnosis system, a disease is further divided into subtypes according to the symptoms, and the intervention would be adjusted accordingly, so the symptom subtype was also extracted, as mentioned in the article. Third, given that the harmful reaction to an intervention is important for the assessment of the safety of this intervention, adverse events were also extracted.

As regular expression requires the establishment of rules according to the content of target fields, target fields containing content with similar characteristics were extracted in a similar manner. The target fields were categorized into three forms: numerical fields, categorical fields, and descriptive fields (Table 1).

The numerical fields primarily contained clinically relevant numbers, which are essential elements in all clinical articles. Considering the extraction of age distribution, when the full text of an article was divided into sections or subsections, the age distribution patterns were extracted from each section/subsection for each of the groups tested separately. To extract this information regarding age distributions, rules were created according to the detailed description patterns. Specifically, keywords for localization that may appear in different patterns such as the different names used to refer to the control group like "placebo group/Western medicine group/Western drug group/conventional group" were standardized by replacing them with those of the "control group." The information was then localized by keywords, and detailed digits were extracted and marked by their locations (i.e., the former digit is the minimum or average age and the latter digit is the maximum age or variance). Subsequently, the unit was assigned to each digit accordingly; if a possible unit was absent, the unit of the latter digit was

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Target fields divided into three categories according to contents

Field category	Topic	target fields
Numerical fields	Age distribution	Minimum age, maximum age, average age, and age variance of total and each group
	Sample sizes	The number of cases for total and each group, and the number of patients who drop out
	Sex distribution	The distribution of male and female cases in total and for each group
	Course of disease	Minimum course, maximum course, average course, and course variance of total and each group
	Course of treatment	Course of treatment
	Observation time	Observation time
	Intervention	Frequency
Categorical fields	Diagnosis	Modern medicine diagnosis, modern medicine diagnostic reference, TCM diagnosis, TCM symptom subtype, and TCM diagnostic reference
	Study design	Random assignment, blinding method, and allocation concealment
	Risk of study design	Risk of random assignment, risk of blinding method, and risk of allocation concealment
	Intervention	Intervention type, and intervention substance/intervention acupoint/physiotherapy type
	Outcome indicator	Outcome indicator
Descriptive fields	Intervention	Intervention on the control group, and Intervention on the experimental group
	Inclusion and exclusion criteria	Inclusion and exclusion criteria
	Keywords	Keywords
	Adverse events	Adverse events

Abbreviation: TCM, traditional Chinese medicine.

assigned to the former digit (Figure 7). Age distributions were typically presented as (1) maximum and minimum; (2) maximum, minimum, and average; (3) maximum, minimum, average, and variance; or (4) average and variance. In each category, most descriptions originated from two to three elementary statements, with some modifications. Rules corresponding to these statements were grouped, for specific articles to be matched. Unmatched rules return null results, only the matched rules return the correct results and are retained. In cases where decimal points were not consistently placed, especially in some articles with low-quality PDF formats, corrections are required. Specifically, if the age was unrealistically large or small, or if the minimum age was larger than the maximum age, they were also marked out for subsequent manual examination. Other target fields were extracted in a similar manner, with the extraction target changing accordingly.

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The categorical fields have limited content. A typical categorical field is the modern medicine diagnosis reference or TCM diagnosis reference. Using the diagnosis, type 2 diabetes, as an example, there are many diagnostic criteria by the World Health Organization (WHO) or based on professional guidelines in documents, such as the *China guideline for type 2 diabetes*. Because the total numbers of diagnostic criteria are limited, TCM experts created a dictionary based on each disease. This study/method utilized a combination of dictionary-based and rule-based methods to extract the target contents precisely. Most categorical fields are extracted similarly to that of the diabetes diagnosis, whereas some target fields such as the intervention substance have enormous possible content. For example, thousands of materials can be used as intervention substances in the TCM intervention and referencing books like *Chinese pharmacopoeia* were used to cover all the possible materials.

The descriptive fields primarily contained declarative events that were often located by keywords from sections/subsections. They contained data that could not be summarized as numbers or words. Thus, sentences describing the fields were extracted directly. Researchers can use numerical and categorical fields to narrow down the search range, and if special needs exist, the descriptive fields can be used to facilitate further information extraction, either by automation technologies or manual methods.

6 | DISCUSSION

In the present study, a precision-preferred comprehensive information extraction system based on rule extraction is introduced. This system fulfills the long-neglected requirements of information extraction for Mandarin TCM clinical articles. The original PDF files of various types and scanned quality were standardized using a hybrid of machine vision and the OCR technique. By systematically dividing the TCM clinical article into sections/subsections and combining rule-based extraction techniques with dictionaries created by TCM experts, this system could provide a precise and comprehensive database for future systematic reviews or meta-analyses, which relieves TCM researchers of the laborious, unscalable, and inefficient manual extraction technique.

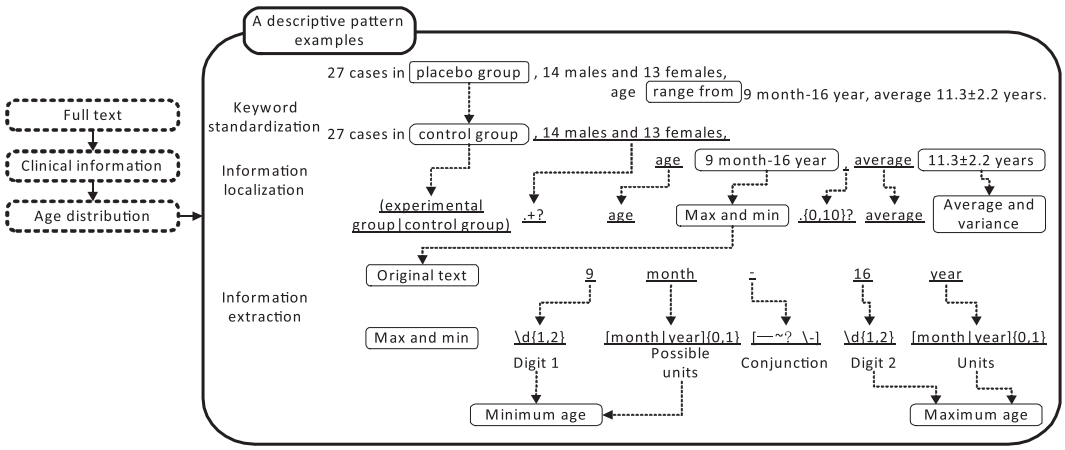


FIGURE 7 Extracting the minimum age and maximum age from a descriptive pattern example

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Prior studies suggested that conducting a systematic review manually required an average of 67 weeks,[23] and a meta-analysis required more than 1000 h of expert manual labor.[24] PubMed search results show that there are over 300,000 results related to systematic reviews or meta-analyses, suggesting a total time investment of approximately 270,000 years. Systematic reviews or meta-analyses in TCM are less frequent, yet there are still thousands of results in CNKI, and it can be predicted that the number will greatly increase in the future. Our system could provide a public foundational data for TCM researchers to save enormous time that would otherwise be devoted to information extraction from TCM clinical articles and reduce repetitive work. With the established comprehensive information extraction system for TCM clinical articles, all useful information was extracted and saved in a structured format, such that future researchers can directly locate information they are interested in from our database, rather than examining and extracting information through original articles.

By manually checking a random sample of information extracted using the method presented in this study, it was suggested that the precision of rule-based extraction can be 96%–100%, but the recall varies greatly from 60% to 95%. Most errors were introduced in identifying articles in low-quality PDF formats. Some target fields were highly irregular, resulting in a recall as low as 60%. Although our system may not be able to cover all relevant articles, it can process more than 5000 articles per hour, and by marking target fields as correct or in need of manual operation, this method can still greatly decrease the cost of the information extraction process.

Rule-based information extraction is a classic method with high precision and low recall,[25] and is an important approach for processing vast amounts of unstructured data. Manual creation of rules requires considerable work by qualified researchers, which greatly increases the cost of rule-based extractions.[26] Hence, rule-based extraction methods are less frequently used, and most information extraction work has been done using machine learning-based technologies to develop low-cost models. However, rule-based information extraction is more efficient for clinical articles compared with its usage in many other fields for the following reasons. First, in contrast to many fields with unlimited changing contents, the contents of clinical articles were relatively limited and fixed. In traditional manual extraction from clinical articles, TCM experts first reached a consensus on the extraction rules. In most instances, these rules were discarded after the completion of the study. In our work, the consensus on extraction rules was created by experts and then generalized as regular expressions that can be easily reused. The reutilization of expert experiences contributes to the consistency of extraction at a relatively low cost. Second, in many research fields, experimental objectives, aims, and methods vary greatly. In contrast, clinical studies are usually designed to perform interventions on people, aiming to evaluate treatment effects, and the results are usually presented in semistructured patterns based on the PICOS model, which makes the digitization procedure more efficient.

Assessment of TCM clinical articles requires systematic information extraction that covers all the elements of the PICOS model rather than a single concept or the relationship between a pair of concepts, which cannot be achieved by previous information extraction studies designed to solve problems in scattering topic areas. With the comprehensive information extraction system established in this study, we evaluated the designs of a subset of TCM clinical studies, and our results revealed that many clinical articles have information deficiencies that may weaken the evidence effect of these articles. Specifically, most information deficiencies exist in the study design section. Many studies do not provide information regarding the blinding methods. Furthermore, vague statements, such as "randomly assigned the patients into

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

different groups," were usually used to describe the allocation concealment. Some authors used methods that are highly likely to introduce bias; for example, they may randomly assign patients into different groups according to registration order/patient number/date. Additionally, it was found that the details on the age or disease course distributions were often missing (e.g., some articles only listed ranges without reporting the average value or variance). Information on adverse reactions was also frequently lacking.

In addition to providing a public foundational data for TCM researchers and helping them identify problems in previous TCM clinical trials, with our comprehensive information extraction system, templates for TCM clinical trials can be set up to guide TCM researchers to present their work properly, which would significantly contribute to the structuration of TCM clinical articles. In our future work, we hope to continue to extend the scope of our focused research themes and improve our rules to increase both precision and recall.

ACKNOWLEDGMENTS

This study was funded by the National Key R&D Program of China (No. 2019YFC1709801/2019YFC1709800) and the State Key Laboratory of Dampness Syndrome of Chinese Medicine (No. SZ2020ZZ09/SZ2021ZZ30).

ORCID

Ye Xia http://orcid.org/0000-0003-0195-4212
Jianxiong Cai http://orcid.org/0000-0001-6811-2242
Yizhen Li http://orcid.org/0000-0003-4710-9049
Zhili Dou http://orcid.org/0000-0001-7436-9626
Yunan Zhang http://orcid.org/0000-0003-3195-8564
Lin Wu http://orcid.org/0000-0002-6653-3059
Zhe Huang http://orcid.org/0000-0002-9846-7295
Shujing Xu http://orcid.org/0000-0002-6643-8123
Jiayi Sun http://orcid.org/0000-0002-7056-1116
Yixing Liu http://orcid.org/0000-0001-5992-8450
Darong Wu http://orcid.org/0000-0003-3159-7359
Dongran Han http://orcid.org/0000-0003-3630-5036

REFERENCES

1. Landhuis E. Scientific literature: information overload. *Nature*. 2016;535(7612):457-458.
2. Barron BA, Bukantz SC. The evaluation of new drugs. Current food and drug administration regulations and statistical aspects of clinical trials. *Arch Intern Med*. 1967;119(6):547-556.
3. Extance A. How AI technology can tame the scientific literature. *Nature*. 2018;561(7722):273-275.
4. Müller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLOS Biol*. 2004;2(11):e309.
5. Tshitoyan V, Dagdelen J, Weston L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*. 2019;571(7763):95-98.
6. Roberts K, Alam T, Bedrick S, et al. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *J Am Med Inf Assoc*. 2020;27(9):1431-1436.
7. Nadeau D, Sekine S. A survey of named entity recognition and classification. *Lingvisticae Invest*. 2007;30:3-26.
8. Yang Y, Xun H, Hong-liang Y. A review of relation extraction *Data Anal Knowl Discovery*. 2013;29:30-39.
9. Blumenthal D. Launching HITECH. *N Engl J Med*. 2010;5:382-385.
10. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inf*. 2018;77:34-49.

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

11. Jensen K, Soguero-Ruíz C, Mikalsen KO, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep*. 2017;7(1):46226.
12. Wang Y, Chen E, Pakhomov S, et al. Automated extraction of substance use information from clinical texts. *AMIA Annu Symp Proc*. 2015;2015:2121-2130.
13. Chikka VR, Mariyasagayam N, Niwa Y, Karlapalemm K. Information extraction from clinical documents: Towards disease/disorder template filling. *Presented at 8th International Conference of the CLEF Association*, Dublin, Ireland, September 2017.
14. Xie J, He J, He W, Hu C, Hu K, Jiang R. Research on structured information extraction method of electronic medical records of traditional chinese medicine. *Presented at IEEE International Conference on BioInformaticsand Biomedicine*, Seoul, South Korea, December 2020.
15. Huang Z, Sun W, Deng H, et al. Data integrity of randomized controlled trial in TCM. *Chin J Evid Based Med*. 2021;21(10):1211-1218.
16. Fang Y, Huang H, Chen H, Juan H. TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. *BMC Complement Altern Med*. 2008;8:58.
17. Zhang H, Ni W, Li J, Zhang J. Artificial intelligence-based traditional Chinese medicine assistive diagnostic system: validation study. *JMIR Med Inf*. 2020;15 8(6):e17608.
18. Chen YB, Tong XF, Ren J, Yu CQ, Cui YL. Current research trends in traditional Chinese medicine formula: a bibliometric review from 2000 to 2016. *Evid Based Complement Alternat Med*. 2019;2019:3961395.
19. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8(1):163.
20. Marshall IJ, Joël K, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inf Assoc*. 2016;1:193-201.
21. Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q. Regular expression-based learning to extract bodyweight values from clinical notes. *J Biomed Inf*. 2015;54:186-190.
22. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inf Assoc*. 2006;13(6):691-695.
23. Borah R, Brown A, Capers PL, Kaiser K. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2):e012545.
24. Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. *J Am Med Assoc*. 1999;282(7):634-635.
25. Tang J, Hong M, Zhang D, Li J. *Emerging Technologies of Text Mining: Techniques and Applications*. IGI Global; 2008 (Chapter 1).
26. Chiticariu L, Li Y, Reiss FR. Rule-based information extraction is dead! Long live rule-based information extraction systems!. *Presented at Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, October 2013.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Xia Y, Cai J, Li Y, et al. A precision-preferred comprehensive information extraction system for clinical articles in traditional Chinese Medicine. *Int J Intell Syst*. 2022;37:4994-5010. doi:10.1002/int.22748

1098111x, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/int.22748 by Mohammad Ali Jinnah University, Wiley Online Library on [03/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License