HHS Public Access

Author manuscript

Stud Health Technol Inform. Author manuscript; available in PMC 2016 July 22.

Published in final edited form as:

Stud Health Technol Inform. 2013 ; 192: 662–666.

Analyzing Differences between Chinese and English Clinical Text: A Cross-Institution Comparison of Discharge Summaries in Two Languages

Yonghui Wu[a,*], **Jianbo Lei**[a,b,*], **Wei-Qi Wei**[c], **Buzhou Tang**[a], **Joshua C. Denny**[c], **S. Trent Rosenbloom**[c], **Randolph A. Miller**[c], **Dario A. Giuse**[c], **Kai Zheng**[d], and **Hua Xu**[a]

[a]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TN, USA

[b]Center for Medical Informatics, Peking University, Beijing, China

[c]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

[d]Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA

Abstract

Worldwide adoption of Electronic Medical Records (EMRs) databases in health care have generated an unprecedented amount of clinical data available electronically. There has been an increasing trend in US and western institutions towards collaborating with China on medical research using EMR data. However, few studies have investigated characteristics of EMR data in China and their differences with the data in US hospitals. As an initial step towards differentiating EMR data in Chinese and US systems, this study attempts to understand system and cultural differences that may exist between Chinese and English clinical documents. We collected inpatient discharge summaries from one Chinese and from three US institutions and manually analyzed three major clinical components in text: medical problems, tests, and treatments. We reported comparison results at the document level and section level and discussed potential reasons for observed differences. Documenting and understanding differences in clinical reports from the US and China EMRs are important for cross-country collaborations. Our study also provided valuable insights for developing natural language processing tools for Chinese clinical text.

Keywords

Cross Language; clinical notes; discharge summary

Introduction

Recently, the Chinese government announced ambitious national health reform plans. It has allocated tremendous funds to improve the health care system in China. For example, a

Address for correspondence: Yonghui Wu, yonghui.wu@uth.tmc.edu.

*These authors contributed equally to this work

recent report indicated that health insurance now covers 95.6% of the population in China [1]. The latter may be one of the greatest healthcare accomplishments worldwide. Health information technology (HIT) stands as one of the eight supporting pillars necessary to achieve Chinese healthcare reform goals. The Chinese government views Electronic Medical Record systems (EMRs) as an essential component for modern hospital management, with the potential to improve the efficiency, quality, and safety of health care. The Chinese Ministry of Health (MOH) has established a standards bureau that in 2009 proposed a series of HIT templates covering EMR basic architectures and data standards [2]. Up to now, many urban hospitals in China adopted and used EMR systems to a variable extent [3]. To accelerate EMR adoption in rural hospitals, the Chinese government allocated 3.9 billion RMB (approximately $600 million US) in 2011 to a pilot program for implementing EMRs in about 200 hospitals [4, 5]. Given the large population of China, the rapid growth in standardized EMR databases there will soon accumulate unprecedented amounts of electronically available clinical data that can support clinical and translational research.

In the US, large academic medical centers have implemented EMR systems for more than three decades and have established large practice-based longitudinal datasets [6]. Recently, the growth of EMRs in US is being fueled by federal legislation that provides generous financial incentives to institutions demonstrating aggressive application and “meaningful use” of comprehensive EMRs [7, 8]. Major efforts are already underway to link these EMRs across US institutions for clinical and translational research. The US EMR databases have been successfully used for various types of studies such as observational comparative effectiveness research [9], genomic [10], and pharmacogenomic studies [11].

Recently, there has been an increasing trend in US and western institutions towards collaborating with China on public health, clinical, and translational research based on EMRs [12, 13]. It is very likely that patient records stored in the EMR systems in China will also become an invaluable asset supporting international collaborative research endeavors. Due to the differences in culture and practice patterns between China and US, EMR data in Chinese hospitals is likely to have different characteristics than data from US institutions. It is important for international collaborations to understand any differences that might exist. Nevertheless, few published studies have compared available EMR data in China versus in the US.

Various EMR systems can contain data in numerous formats, including as both structured and unstructured information. For example, EMR systems typically store narrative clinical reports, containing detailed treatment and outcome information for individual patients. Such reports comprise highly valuable resources for clinical research.

As an initial step towards differentiating EMR data in Chinese and US systems, this study attempts to understand system and cultural differences that exist between Chinese and English clinical documents. More specifically, the study collected in-patient discharge summaries from one Chinese and from three US institutions and investigators manually analyzed three major clinical EMR components: medical problems, tests, and treatments. We report comparison results at the document level and section level.

Methods

Data sets

Organizers of the 2010 i2b2 (Center of Informatics for Integrating Biology and the Bedside) clinical NLP challenge [14], collected 826 clinical notes, of which 646 are inpatient discharge summaries, from three US hospitals: University of Pittsburgh Medical Center (UPMC), Partners Healthcare (PARTNERS), and Beth Israel Deaconess Medical Center (BETH). For each clinical note in the collection, domain experts manually annotated three clinically important components: medical problems (e.g., diseases and symptoms), tests (e.g., lab tests), and treatments (e.g., medications and procedures), by following annotation guidelines developed by i2b2 challenge organizers. [15]

This study included and analyzed all 646 2010 i2b2-related discharge summaries and compared them to Chinese clinical notes. We collected one-month (March 2011) of discharge summaries from the EMR database of Peking Union Medical College Hospital (PUMCH) in China. Following guidelines similar to those used in the 2010 i2b2 NLP challenge, we developed an annotated corpus of these Chinese discharge summaries for use in this study. After excluding very short notes (incomplete notes), we randomly selected 400 discharge summaries from the PUMCH pool for this study. All patient identifiers in the notes were manually removed by PUMCH physicians before the notes were sent to researchers for annotation. Using transliterated i2b2 guidelines two native Chinese-speaking domain experts manually annotated problems, tests, and treatments in each note. To calculate the inter-rater agreement for annotation, 40 notes were identical for the two annotators. These 400 annotated discharge summaries in Chinese were used and compared with the 646 discharge summaries in English in this study.

Analytic Methods

We conducted content analysis on the 646 English and 400 Chinese discharge summaries using Charmaz's grounded theory approach [16]. We approached the data with no prior assumptions and generated descriptive statistics based on the content of the notes. We analyzed the data with a focus on understanding the distributions of three types of important *clinical entities* (Problems, Tests, and Treatments) at both document and section levels, as well as the differences of such distributions between Chinese and English clinical text.

Document level analysis—At the document level, we conducted two experiments: (1) compare the vocabulary distribution and the density of clinical entities (defined as the average number of clinical entities in each document) in Chinese and English corpora; and (2) report relative frequency of three types of entities for each institution. Zipf's distribution is widely used to describe the vocabulary frequency by plotting a log-scale graph between frequency and rank. We collected all the words from the two corpora and then ranked the words according to their frequencies to present the curve in log scale. As there are no spaces denoting word breaks in the Chinese corpus, the Stanford Word Segmenter [17] trained on Penn Chinese Treebank corpus [18] was used to identify individual Chinese words. In experiment 2, the relative frequency for a specific entity type is defined as the number of entities belong to this type divided by the total number of all three types of entities. We

calculated the relative frequencies of three different entity types: Problem, Test, and Treatment for all four institutions.

Section level analysis—At the section level, we focused on measuring the density of clinical entities (defined as the average number of clinical entities for a given section) and the differences of entity density among four institutions for different sections. Section identification in clinical text is not a trivial task [19]. In this study, we developed an ad-hoc approach to identify sections in Chinese and English notes.

- Detect candidate section headers -- a program was developed to detect all the candidate section headers using the colon, upper case letter and other features.
- Group section headers -- we manually reviewed all the candidate sections to remove false positives and group all the variations according to the contents under section header.
- Match section headers -- two domain experts (authors WW -- who is familiar with both Chinese clinical notes and English clinical notes, JD -- a domain expert in English clinical notes) working together to match the corresponding section headers between English corpus and Chinese corpus according to the content under each section.

Once sections were identified, we reported the average number of clinical entities for each section. To further understand the differences in section content, we also compared the average number of entities within each section in both the English and Chinese corpora.

Results

Based on the annotation results of 40 overlapped discharge notes from PUMCH corpus, the token level Kappa score between the two annotators was 0.99.

Figure 1 shows the word frequency distribution for English corpus and Chinese corpus, showing a typical distribution for Zipf's law. As the English corpus contains more notes than Chinese corpus, the curve for English is above the Chinese corpus (labeled as PUMCH). Figure 2 shows the normalized distribution of entities in English corpus and Chinese corpus. The curve for English corpus descends smoothly, whereas, the curve for PUMCH ends with a sharp decrease, indicating that the English corpus appeared to use a more diverse vocabulary; however, such analysis is complicated by the differences in word form variation between the two languages.

Table 1 shows the number of different types of entities across four different institutions. Compared with the three US institutions, the PUMCH corpus had fewer Treatment entities than the English corpora. The relative frequencies of the three types of entities within each individual institute are shown in Figure 3. The relative frequencies are different among the four institutions, with the unique traits of PUMCH compared to the three English institutions more obvious. PUMCH had a higher proportion of 'Problem' entities and fewer 'Treatment' entities than in English institutions.

After grouping the variations and matching the section headers between Chinese corpus and English corpus, two domain experts detected 12 common, high-level sections appearing in both English and Chinese corpora. In this study, we focused on the comparison of 9 common sections appearing in at least 10 notes. Table 2 shows the density of entities within the 9 sections across four institutions. The results show that the density of entities is markedly different between PUMCH corpus and the English corpora, where the minimum density in English corpora is at least twice of PUMCH corpus in the following three sections: PS, DM, and DI.

Discussion

This study compared the distribution of three types of important clinical entities (i.e., problems, tests, and treatments) in inpatient discharge summaries among three US institutions and one Chinese institution. Understanding such structural differences may help to maximize the value of EMR data acquired in Chinese hospitals when the data are utilized for secondary use purposes such as international collaborations on clinical, translational, and global health research. These structural differences in clinical documentation may also reflect more fundamental system and cultural differences in patient care delivery in China vs. that in US. This knowledge can be critical to the success of collaborative research efforts between the two countries, and between China and other western countries more broadly

The study revealed some interesting data and differences. First, the number of clinical entities per document varied widely among different institutions, even for three US institutions (e.g., 60.30 for UPMC vs. 149.57 for BETH). Further investigation should examine potential explanations for this variability – for example, the effects of clinical documentation methods at different institutions (e.g., directly typed in vs. dictated and transcribed notes). Of note, the Chinese discharge summaries contained fewer Treatment clinical entities than any US institution’s discharge summaries. Again, further investigation should determine why this difference exists, e.g., whether physician workloads varied between settings. Whether it indicates that fewer procedures and medications are ordered in clinical practice in China is not certain; but it is interesting and worth conducting further investigation. Other potential causes for the greater content in US include 1) billing requirements and a 2) a more complex US medicolegal environment in which more thorough testing and discussion of problems may be performed in order to provide defense against a perceived higher risk of litigation.

When analyzing clinical term distributions within different document sections, we noticed that some frequent sections in English discharge summaries, such as “Current Medications” and “Social History”, were not found in Chinese notes. Manual review by a Chinese physician showed that this information could be scattered among different sections. For example, medication information could be recorded in a patient’s Past Medical History section, e.g., "the patient was diagnosed with HTN in 1995. She is taking a beta blocker (Metoprolol) and her BP is normal”. This also may explain the differences between US and Chinese notes in entity frequency distribution for a given section (Table 2). Chinese physicians in the team (JL and WQ) thought this was an important finding, as it provides

valuable information about how to re-organize the structure of Chinese clinical notes for better representation and communication of patient information.

One of the challenges of using EMR data for medical research, which exists for both US and Chinese EMRs, is that much of the detailed clinical information is embedded in narrative clinical reports, which are not directly usable for analysis. Much effort has been devoted to develop natural language processing (NLP) technologies for English clinical text [20, 21, 22] and some approaches have shown limited success [23]. However, little work has been done on NLP regarding Chinese clinical text in EMRs. This study also provides potential insights relevant to the development of NLP tools for Chinese clinical text. During the vocabulary distribution analysis (Figure 1), we explored the word segmentation methods for Chinese clinical corpus. Different from English, Chinese text do not have spaces between words, which makes it more difficult for identifying word boundaries. Our initial analysis showed that clinical dictionary resources helped in word segmentation of Chinese clinical text. In addition, the section analysis of Chinese clinical text is also helpful for NLP research. Further studies on Chinese clinical text processing are one area for future work.

This study has limitations. One of the major limitations was that the analysis of Chinese clinical text was conducted on notes from one institution in China only. Therefore the results regarding Chinese notes might not be representative. Future studies should include Chinese clinical notes from multiple institutions in China. Another limitation was that i2b2 notes lacked information about the clinical settings in which the notes were generated. Additional investigation on healthcare and documentation processes would also provide useful explanations for the differences.

Documenting and understanding system/cultural differences in EMR documents from the US and China are important. These differences may reflect fundamental differences in patient care delivery, and the different structures of healthcare systems. Mastering the differences will be critical in helping US/western researchers understand how to properly interpret and computationally reuse clinical documents produced in either healthcare system relative to the other. In addition, such learning may also inform opportunities to develop novel NLP tools for processing narrative documents in Chinese, or fine-tune tools that were originally developed in the English context.

Acknowledgments

This study was supported by grant from the NLM R01LM010681. We would like to thank the 2010 i2b2/VA challenge organizers for the development of the English corpora used in this study.

References

1. Lim MK, Yang H, Zhang T, Feng W, Zhou Z. Public perceptions of private health care in socialist China. Health Aff (Millwood). 2004; 23(6):222–34. [PubMed: 15537602]
2. EMR basic architecture and data standards. http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohbgt/s6694/200908/42155.htm
3. Report of 2011–2015 Market Survey and prediction of development of China's EMR. http://www.chinairr.org/report/R10/R1006/201110/31-86054.html

4. Chinese MOH Notice on First 97 trial hospitals for EMR. http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohyzs/s3586/201105/51779.htm
5. Chinese MOH Notice on Second 92 trial hospitals for EMR. http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohyzs/s3586/201111/53273.htm
6. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified dna biobank to enable personalized medicine. Clin Pharmacol Ther. 2008; 84(3):362–9. [PubMed: 18500243]
7. Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. N Engl J Med. 362(3):192–5. [PubMed: 20089969]
8. Secretary Sebelius Announces Final Rules To Support 'Meaningful Use' of Electronic Health Records. US Department of Health and Human Service;
9. Pace WD, et al. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009; 151(5):338–40. [PubMed: 19638402]
10. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struewing JP, Wolf WA. eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011; (4):13. [PubMed: 21269473]
11. Xu H, Jiang M, Oetjens M, Bowton EA, Ramirez AH, Jeff JM, Basford MA, Pulley JM, Cowan JD, Wang X, Ritchie MD, Masys DR, Roden DM, Crawford DC, Denny JC. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. J Am Med Inform Assoc. 2011; 18(4):387–91. DOI: 10.1136/amiajnl-2011-000208 [PubMed: 21672908]
12. Wang X, Wang E, Marincola FM. Translational Medicine is developing in China: A new venue for collaboration. J Transl Med. 2011 Jan 4.9:3. [PubMed: 21205297]
13. Zhou, Jiebai; Wu, Duojiao; Liu, Xinqing, et al. Translational medicine as a permanent glue and force of clinical medicine and public health: Perspectives (1) from 2012 Sino-American symposium on clinical and translational medicine. Clin Transl Med. 2012; 1:21. [PubMed: 23369646]
14. Uzuner, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011; 18(5):552–6. [PubMed: 21685143]
15. I2b2 2012 annotation guidelines. https://www.i2b2.org/NLP/Relations/assets/Assertion%20Annotation%20Guideline.pdf
16. Charmaz, K. Constructing grounded theory. SAGE; 2006.
17. Chang, Pi-Chuan; Galley, Michel; Manning, Christopher D. Proceedings of the Third Workshop on Statistical Machine Translation (StatMT '08). Association for Computational Linguistics; Stroudsburg, PA, USA: 2008. Optimizing Chinese word segmentation for machine translation performance; p. 224-232.
18. Xue, Naiwen; Xia, Fei; Chiou, Fu-dong; Palmer, Marta. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Nat Lang Eng. Jun; 2005 11(2):207–238.
19. Denny JC, Spickard A 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc. 2009 Nov-Dec;16(6):806–15. [PubMed: 19717800]
20. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association : JAMIA. Mar-Apr;1994 1(2):161–174. [PubMed: 7719797]
21. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17–21. [PubMed: 11825149]
22. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. Sep-Oct;2010 17(5):507–513. [PubMed: 20819853]
23. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128–144. [PubMed: 18660887]

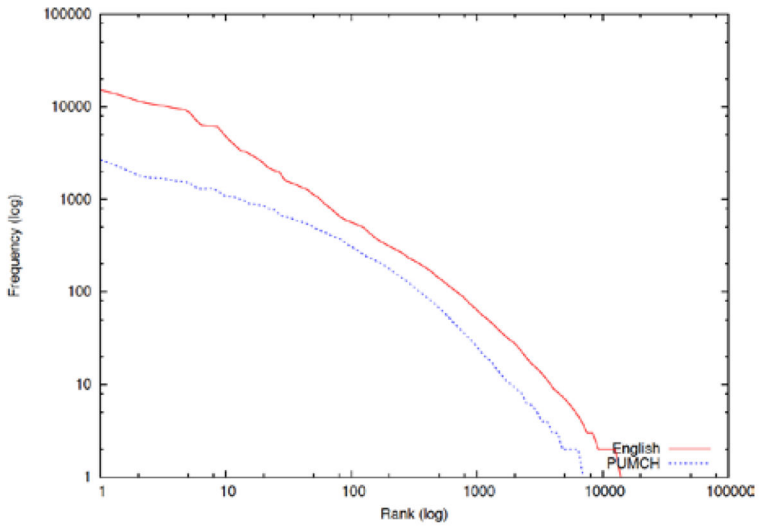

Figure 1.
Zipf's distribution of vocabularies

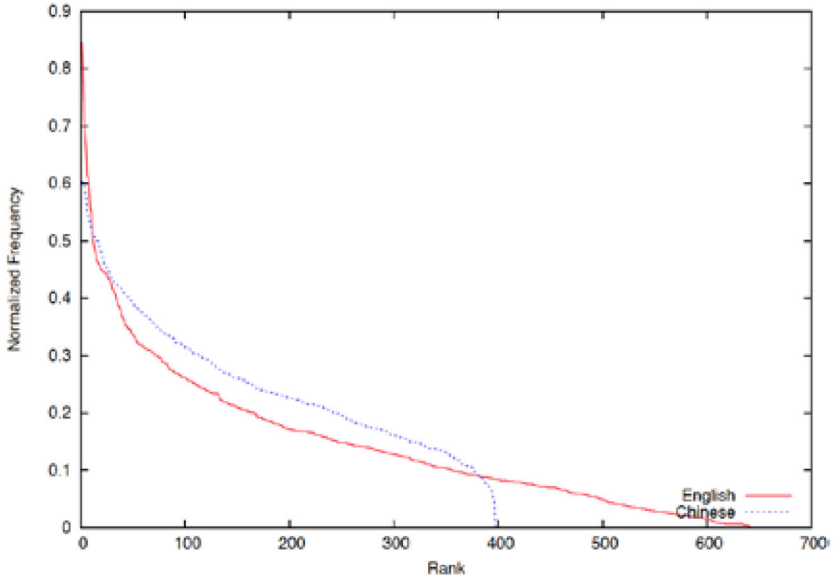

Figure 2.
Normalized distribution of annotated entities

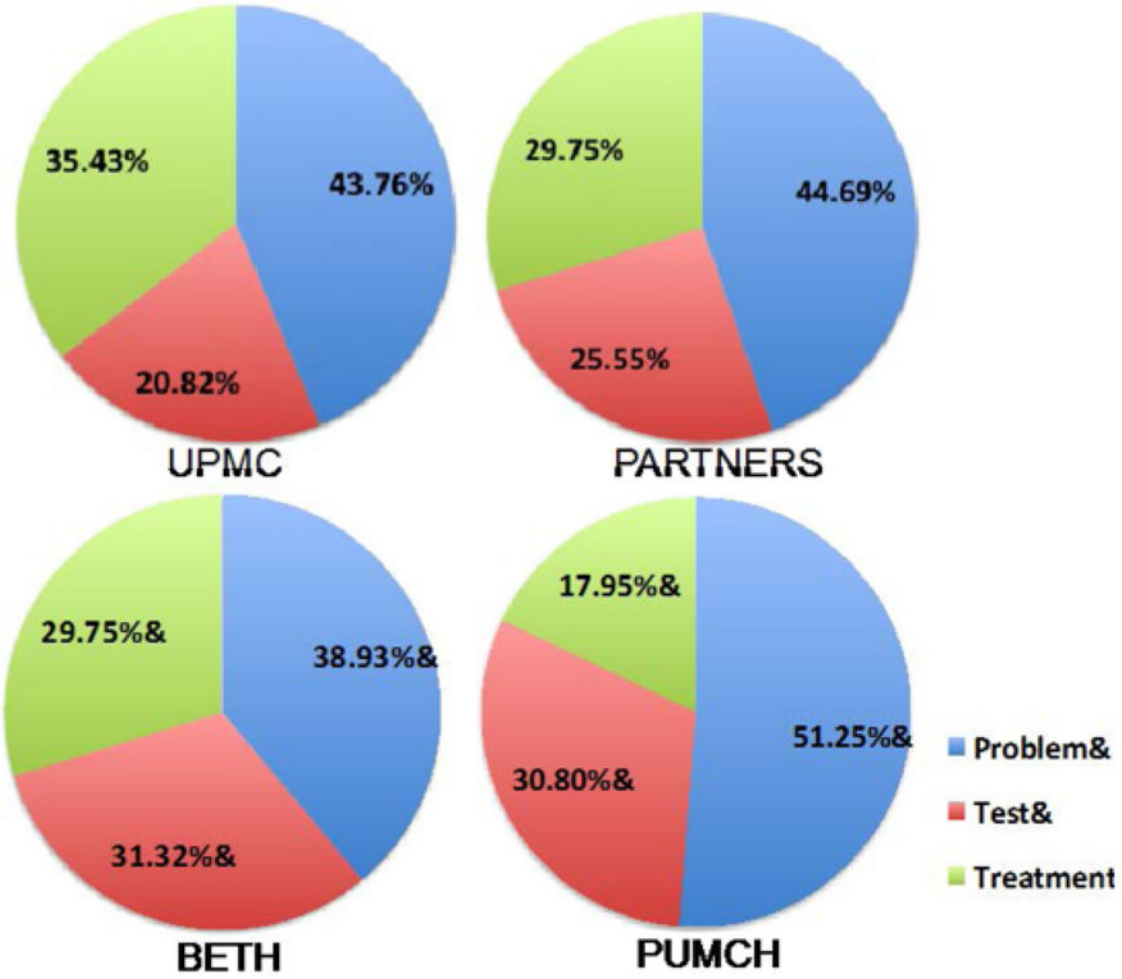

Figure 3.
Relative frequency of Problem, Tests, and Treatments in three English institution: UPMC, PARTNERS, and BETH, and one Chinese institution: PUMCH

Table 1

Distribution of different types of entities

Corpus	# of Doc	Type	# of Entity	Average # of entity per note	Relative Frequency
UPMC (English)	220	Prob	5805	26.39	43.76%
		Test	2762	12.55	20.82%
		Treat	4700	21.36	35.43%
		All	13267	**60.30**	--
PARTNERS (English)	235	Prob	8542	36.35	44.69%
		Test	4884	20.78	25.55%
		Treat	5686	24.20	29.75%
		All	19112	**81.33**	--
BETH (English)	191	Prob	11122	58.23	38.93%
		Test	8947	46.84	31.32%
		Treat	8499	44.50	29.75%
		All	28568	**149.57**	--
PUMCH (Chinese)	400	Prob	20159	50.40	51.25%
		Test	12114	30.29	**30.80%**
		Treat	7061	17.65	**17.95%**
		All	39334	**98.34**	--

Prob -- Problem, Treat -- Treatment

Stud Health Technol Inform. Author manuscript; available in PMC 2016 July 22.

Table 2

Distribution of entities within matched sections

	UPMCD (English)			PARTNERS (English)			BETH (English)			PUMCH (Chinese)		
Section	**Doc**	**Entity**	**Ave**	**Doc**	**Entity**	**Ave**	**Doc**	**Entity**	**Ave**	**Doc**	**Entity**	**Ave**
PS	131	4453	**33.99**	174	5259	**30.22**	151	6211	**41.13**	389	1761	**4.53**
DM	95	1224	**12.88**	138	1113	**8.07**	123	1418	**11.53**	196	518	**2.64**
DI	47	314	**6.68**	54	271	**5.02**	100	713	**7.13**	168	496	**2.95**
CC	33	377	**11.42**	34	67	**1.97**	77	127	**1.65**	398	960	**2.41**
DD	105	1005	**9.57**	35	126	**3.60**	136	793	**5.83**	387	2742	**7.09**
HOPI	30	486	**16.20**	151	3481	**23.05**	159	4612	**29.01**	398	14713	**36.97**
PE	25	479	**19.16**	142	2489	**17.53**	157	3039	**19.36**	265	4000	**15.09**
PMH	59	659	**11.17**	140	2209	**15.78**	166	3812	**22.96**	220	2365	**10.75**
PL	48	187	**3.90**	41	237	**5.78**	35	699	**19.97**	397	1800	**4.53**

Stud Health Technol Inform. Author manuscript; available in PMC 2016 July 22.