

Knowledge discovery for the treatment of bacteria affecting the liver

Pınar YILDIRIM¹, Kağan ÇEKEN², Osman SAKA³

Aim: Biomedical information is buried in millions of published articles, and so it is necessary to use text mining techniques to skim published articles for relevant information. In this study, we used biomedical text mining techniques to introduce a liver bacterial infection knowledge-acquisition information system.

Materials and methods: Bacteria names were selected from Medline MeSH data and it was searched to identify the most frequent bacteria associated with the liver using a text mining system and time based analyses were used to show the evolution of treatments.

Results: Liver infections constitute a major threat to public health, and our study shows that there is a need for better drugs.

Conclusion: Both pharmaceutical industry and healthcare providers are encouraged to investigate challenges related with major liver infections and create strategies to develop new drugs.

Key words: Biomedical text mining, knowledge discovery, bacteria, liver

Karaciğeri etkileyen bakterilerin tedavisi için bilgi keşfi

Amaç: Biyomedikal bilgiler yayınlanmış milyonlarca makalede gömülüdürler ve bu bilgileri sorgulamak için metin madenciliği yöntemlerinden yararlanılabilir. Biz bu çalışmada, karaciğer bakteriyel enfeksiyonlarına ait bilgi keşfi yapmak amacıyla ile metin madenciliği yöntemlerini kullandık.

Yöntem ve gereç: Bakteri isimleri Medline MeSH data'dan alınmış ve Medline karaciğerle ilişkili en yüksek frekanslı bakterileri bulmak için taranmıştır. Bakteriler seçildikten sonra, herbir bakteri için ilişkili makaleler PubMed'den taranmış, metin madenciliği yöntemleri ve tedavilerin zamana bağlı değişimlerini gösteren analizler kullanılarak ilişkili tedaviler belirlenmiştir.

Bulgular: Karaciğer enfeksiyonları toplum sağlığı için önemli bir tehlike oluşturmaktadırlar ve bizim çalışmamız daha iyi ilaçlara ihtiyaç olduğunu göstermiştir.

Sonuç: Hem ilaç endüstrisi hem de sağlık hizmeti sağlayıcıları önemli karaciğer enfeksiyonlarındaki zorlukları araştırmak ve yeni ilaçlar geliştirmek için stratejiler yaratmak için teşvik edilmelidirler.

Anahtar sözcükler: Biyomedikal metin madenciliği, bilgi keşfi, bakteriler, karaciğer

Received: 02.09.2010 – Accepted: 02.11.2010

¹ Department of Medical Informatics, Informatics Institute, Middle East Technical University, Ankara - TURKEY

² Department of Radiology, Faculty of Medicine, Akdeniz University, Antalya - TURKEY

³ Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University, Antalya - TURKEY

Correspondence: Pınar YILDIRIM, Sağlık Bilişimi Bölümü, Enformatik Enstitüsü, Orta Doğu Teknik Üniversitesi, İnönü Bulvarı, 06531, Ankara - TURKEY

E-mail: pinar@cankaya.edu.tr

Introduction

The biomedical literature, such as Medline articles, is a rich resource for discovering and keeping up with medical knowledge. For example, information with regard to which drugs are used with a particular disease or changes in drug usage over time is valuable but unfortunately buried in thousands of articles (1). Biomedical text mining techniques play an important role in acquiring knowledge from these articles and they have been applied to numerous studies so far (1).

Bacterial infections can lead to serious life threatening complications and death (2). Bacteria are especially harmful on the liver, which supports almost every organ and is thus vital for survival. Because of its strategic location and multidimensional functions, the liver is also prone to many diseases. In this study, we developed a knowledge discovery method on the treatment of bacteria affecting the liver and we hope that our study provides valuable information to all scientists and experts working on bacteria affecting the efficiency of the liver's mechanism.

We think that our study will also make substantial contributions in the area of medical research. It helps physicians to get the facts embedded in medical articles and to interpret them in order to build

up new medical knowledge. Using our method, pharmaceutical researchers and antibacterial drug designers can backtrack and evaluate drug usage against harmful liver bacteria and develop new effective strategies providing better treatment efficacy.

Materials and methods

In this study, named entity recognition and time based analysis were used to analyse Medline articles. The following steps were applied to extract information about the treatment of harmful liver bacteria (Figure 1):

- Searching Medline articles to find bacteria associated with liver diseases most frequently
- Selecting top ranked bacteria
- Collecting articles for each bacterium from Medline for specific time periods
- Extracting drugs from the published articles
- Normalisation of names of the drugs to obtain drug search standardisation
- Calculating drugs' frequencies and showing differences in liver drug discovery during different time periods

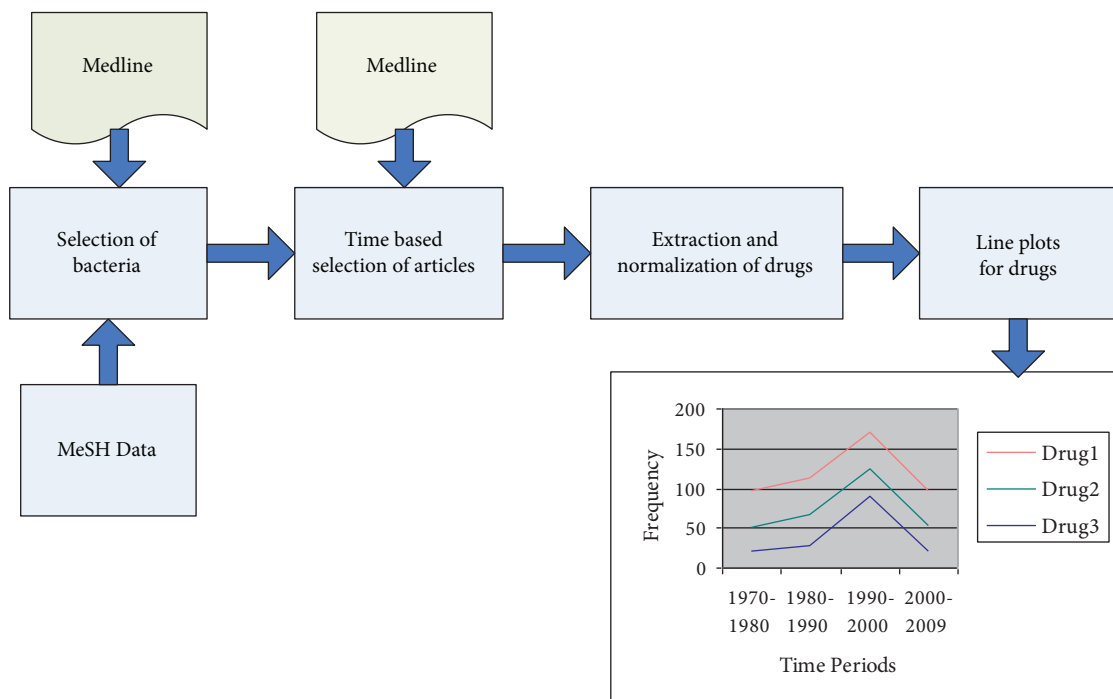


Figure 1. Summary of the research procedure followed in this study.

Bacteria names were selected from MeSH data in the National Library of Medicine (NLM) and Medline was searched to find bacteria that are associated with the liver most frequently. Four bacteria that were encountered in the highest number of articles with the liver in Medline were selected, namely, *Salmonella typhimurium*, *Staphylococcus aureus*, *Mycobacterium tuberculosis*, and *Helicobacter pylori*.

Medline is a collection of biomedical documents and administered by the National Center for Biotechnology Information (NCBI) of the United States National Library of Medicine (NLM) (3). The documents are available on PubMed website. PubMed is a service of the National Library of Medicine and includes over 18 million bibliographic citations from Medline and other life science journals for biomedical articles dating back to the 1950s. Full texts of the articles are not stored; rather, links to providers' sites are provided (4). These links lead to full-text versions of the articles.

Medline abstracts are in XML format and they contain logical markup to organise meta-information. For example, the XML structure of Medline abstracts include meta-information attached to the original document, such as the name of the journal, author list, affiliations, and publication dates as well as annotations inserted by the NLM, such as creation date of the Medline entry and list of chemicals associated with the document, as well as related MeSH headings (5).

After selecting the bacteria, we provided a drug time based analysis for each bacterium in addition to relevant articles, collected from PubMed, and specific drug usage time periods (e.g., 1970-1980, 1980-1990, 1990-2000, and 2000-2009). PubMed offers many search options for users. For example, in order to find articles published in the period of 1970-1980 and relevant to *S. typhimurium* and the liver, a researcher can use the time range option by selecting the "1970-1980" date range. Table 1 shows the obtained total number of articles for the selected bacteria. The following query was used for finding published articles for a considered time period. This query was modified for other time periods to collect time specific articles.

"Salmonella typhimurium "[All Fields] AND "liver"[All Fields] AND ("1970"[EDAT]: "1980"[EDAT])

After retrieving the articles in specific time periods, drug names were found by using a drug filter server that tags drug names from Drugbank (6). The Drugbank database is a bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological, and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information (7).

Table 1. Number of articles for the bacteria selected.

Bacteria	Time Periods				Total
	1970-1980	1980-1990	1990-2000	2000-2009	
	Number of Articles				
<i>Salmonella typhimurium</i>	668	1710	1063	584	4025
<i>Staphylococcus aureus</i>	89	291	304	335	1019
<i>Mycobacterium tuberculosis</i>	174	129	318	350	971
<i>Helicobacter pylori</i>	0	8	186	440	634

A text mining solution, known as the filter server, developed by Rebholz Group at the EBI (European Bioinformatics Institute) was used to extract drug names from Medline articles. The architecture of the biomedical text mining system in EBI is based on regular expressions and provides a framework for the extraction of facts from the biomedical literature. The Monq java is the java class library that binds regular expressions to actions that are automatically executed whenever a match occurs in the text stream being processed. In the case of a match, the associated action can modify the stream or leave it unchanged. Commonly, XML tags are used to mark named entity and other regions of interest in the text. Several thousand regular expression/action pairs can be combined into one machinery, called a Deterministic Finite Automaton (DFA), and can be used in parallel with different computer applications (4,5).

In EBI architecture, there are some filter server solutions and each filter server specialises in recognising the vocabulary of a particular terminology and performing specialised actions depending on the input it receives. A filter server annotates streams of text. The server runs its embedded software on the incoming text to recognise and tag the terminology with XML tags (5). In this study, after using the drug filter server, which tags drug names taken from the Drugbank database, all drugs in the articles are annotated in XML. At the next step, frequencies of drugs were automatically calculated for each time

period. The frequency of a drug provides the number of times a drug appears in the selected articles. A software program that we developed in Java was used to find the frequencies. In addition, drugs' therapy classes were searched in Drugbank to find their categories. After finding categories of all drugs, antibacterial drugs, which are the main class of treatment for bacteria, were selected for time based analysis. Since the number of articles varies in each time period, the frequencies of drugs were divided by the number of articles to get the frequencies. Minitab statistical software was used to create line plots for each bacterium and the differences between frequencies are presented in these plots.

Line plots provide drug time analysis for clinicians to identify drug usage over time and to compare drugs according to their frequencies. Drugs have some variations, such as synonyms and brand names. For example, theophyllin is a synonym of theophylline. On the other hand, Serax is a brand name of oxazepam. Drugbank was searched for each drug, for synonyms and brand names. Drugbank is one of the biggest resources for drugs and currently contains >4100 drug entries, corresponding to >12,000 different trade names and synonyms (6,7). After finding the variations, these names were manually normalised to one specific name. Table 2 shows drug name variations, normalised names, and Drugbank ID, which identifies uniquely each drug in the Drugbank database.

Table 2. Normalised names of some drugs used against bacteria.

Drug Name Variations	Normalised Names	Drugbank ID
Theophyllinse, Theophyllin	Theophylline	DB00277
Erythromycin, Erythrocin Stearate	Erythromycin	DB00199
Ampicillin, Polycillin, Principen	Ampicillin	DB00415
Zidovudine, Retrovir	Zidovudine	DB00495

Results

Salmonella typhimurium

Infections with nontyphoidal *Salmonella* have increased during the last 3-4 decades. Although a decrease in infections has been reported over the last decade, *Salmonella* infections continue to be a major public health concern in many countries and the resistance to antimicrobial drugs appear to pose a particular health risk. In the 1990s, resistance against the following drugs was reported: ampicillin, chloramphenicol, streptomycin, sulphonamides, and tetracycline (8,9). Threlfall et al. investigated the changes in antimicrobial resistance in *S. enteritidis* and *S. typhimurium* from human infection in England and Wales in 2000, 2002, and 2004 and reported that the incidence of resistance to nalidixic acid, coupled with decreased susceptibility to ciprofloxacin, has more than doubled between 2000 and 2004. Resistance to nalidixic acid and ciprofloxacin has changed from 43% in 2000 to 76% in 2004. The occurrence of resistance to ampicillin increased from 5% in 2000 to 8% in 2004, but resistance to tetracyclines and trimethoprim changed very little over the 5-year period (10).

Chen et al. analysed the trend of drug resistance of *Salmonella typhimurium* in Taiwan in 1991-2001. Their study showed that the drug resistance rate for a single drug was the highest for streptomycin at 84.2%, followed by tetracycline at 82.5%, chloramphenicol at 71.9%, ampicillin at 70.2%, and nalidixic acid at 18.4%. Changes in drug resistance of *Salmonella typhimurium* in recent years were studied using new generation antibiotics, such as ciprofloxacin and ceftriaxone. According to their results, ciprofloxacin of the floroquinolones group showed 3.8% drug resistance (9).

Figure 2 shows the line plot of antibacterial drugs for *Salmonella typhimurium*. According to the figure, the frequencies of ampicillin and tetracycline have decreased after the 1980-1990 time period. It also shows that resistance to these drugs was seen in these time periods. However, in the 1990-2000 time period, resistance to tetracycline increased again. Gentamicin and sulphonamides have also decreasing frequencies after 1990-2000. Surprisingly, norfloxacin has an increasing frequency in the 1990s.

Norfloxacin has been proved to be a very broad spectrum anti-bacterial drug and it is one of the new 4-quinolone anti-bacterial agents introduced in 1984 in the world market. Subsequent to norfloxacin, 4 more new quinolone compounds also have come onto the market. Still norfloxacin and its successor ciprofloxacin are able to hold their own market in their clinical use as popular agents for urinary tract infections (11).

Staphylococcus aureus

The first cases of methicillin-resistant *Staphylococcus aureus* (MRSA) infections were reported from the UK in 1961 and MRSA became a major problem in hospital settings worldwide in the 1980s. Although community-acquired MRSA are emerging worldwide, vancomycin-resistant *S. aureus* remain extremely rare. Up until 2007, 3 vancomycin-resistant *S. aureus* cases were reported from the US in 2002 and 2004. Linezolid is one of the new active agents and it is active against MRSA (12,13). Figure 3 shows the line plot of antibacterial drugs for *S. aureus*. Despite their resistance, the frequencies of both methicillin and vancomycin have increased in all time periods. Ciprofloxacin has a decrease after 1990-2000. Rifampin has been increasing but still has a low frequency. Furthermore, linezolid has been seen after the 1990-2000 time period and it has an increasing frequency.

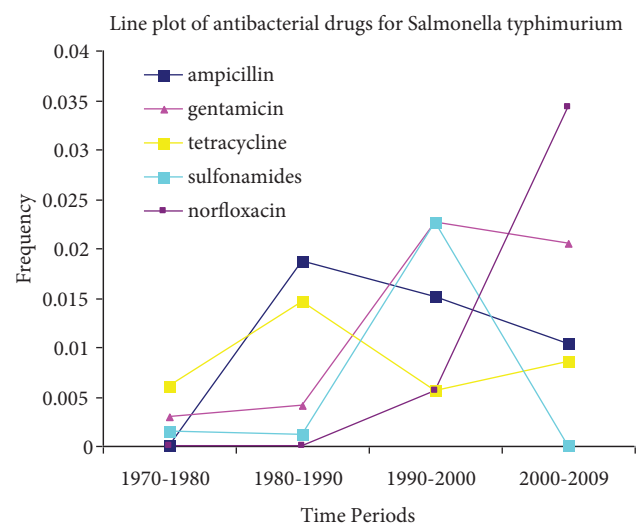


Figure 2. Line plot of antibacterial drugs for *Salmonella typhimurium*.

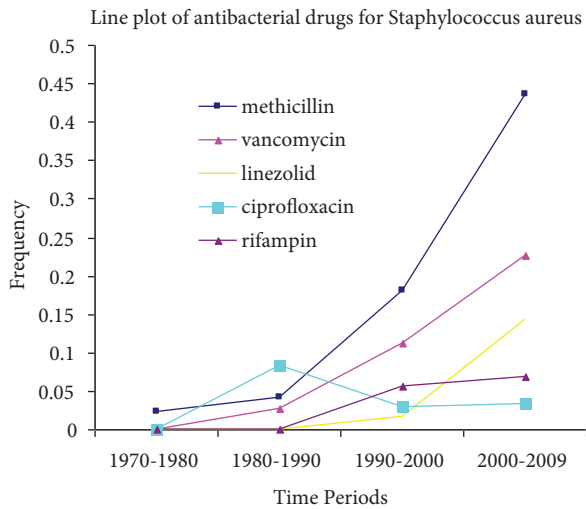


Figure 3. Line plot of antibacterial drugs for *Staphylococcus aureus*.

Mycobacterium tuberculosis

Tuberculosis is caused by *Mycobacterium tuberculosis* and kills approximately 2 million people each year. Due to the intrinsic resistance of *M. tuberculosis* to many antibiotics, chemotherapy of tuberculosis is restricted to a very limited number of drugs, which have to be used in combination for at least 6 months (14).

Multidrug-resistant tuberculosis (MDR TB) is a form of tuberculosis that is resistant to some of the first-line drugs used for the treatment of the disease. It is associated both with a higher incidence of treatment failures and of disease recurrence, as well as with higher mortality than forms of tuberculosis sensitive to first-line drugs. Levofloxacin (LFX) represents one of the few second-line drugs recently introduced in the therapeutic regimens for MDR TB (15).

Rifampicin (RFP) was developed as one of the anti-tuberculosis drugs in 1966 and has been in use since then. Establishment of combination therapy using RFP has been contributing to the treatment/eradication of tuberculosis. A number of rifamycin derivatives, as post RFPs, have been synthesised/developed over the years. Chemical modification of rifamycins has largely been concentrated on the moiety of naphthalene ring because modification of the ansa chain moiety reduces the activity. In 1992, rifabutin was approved as a preventive drug for MAC

infection in AIDS patients in the United States and in European countries (16).

Hsueh et al. summarised data from 1990-2002 in Taiwan and results showed that primary resistance ranged from 4.7% to 12% for isoniazid, 0.7% to 5.9% for rifampin, 1% to 6% for ethambutol, and 4% to 11% for streptomycin (17).

Figure 4 shows the line plot of antibacterial drugs that we obtained for *M. tuberculosis*. According to the figure, in both 1980-1990 and 1990-2000 time periods, significant changes have been observed for some drugs, such as clarithromycin, amikacin, and rifabutin. In addition, streptomycin and ofloxacin have also both increases and decreases in these time periods but these changes are not as sharp as the others.

Helicobacter pylori

The discovery that *Helicobacter pylori* infection is the main cause of most gastroduodenal diseases has been a major breakthrough in gastroenterology. It has dramatically changed the management of these diseases, which are now considered as infectious diseases and are treated with antibiotics (18).

Triple therapy, including 2 antibiotics (amoxicillin and clarithromycin), and a proton pump inhibitor administered for a week has been recommended as the treatment of choice at several consensus conferences. However, this treatment may fail for several reasons, as reported elsewhere. In fact, the main reason for

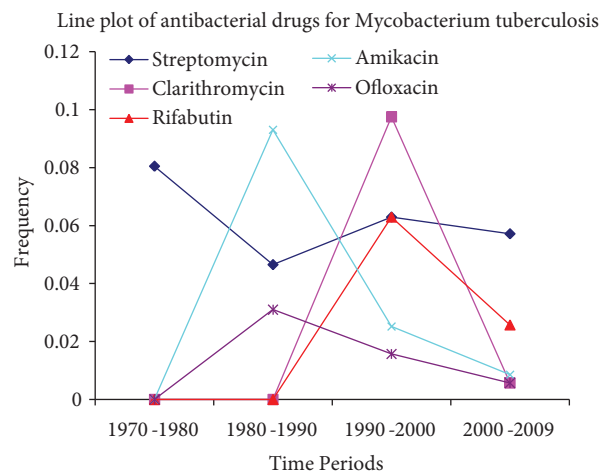


Figure 4. Line plot of antibacterial drugs for *Mycobacterium tuberculosis*.

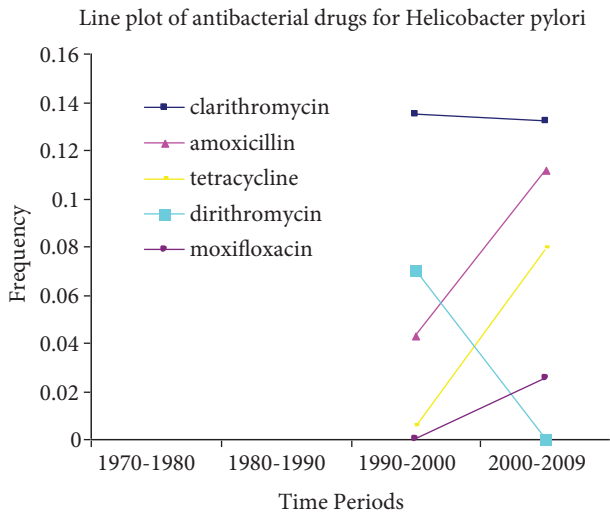


Figure 5. Line plot of antibacterial drugs for *Helicobacter pylori*.

failure was found to be *H. pylori* resistance to one of the antibiotics used (that is, clarithromycin). Other treatments have also been proposed, including metronidazole, fluoroquinolones, and rifamycins for which resistance has become an emerging issue (18).

Figure 5 shows the line plot of antibacterial drugs that we obtained for *H. pylori*. Except clarithromycin and dirithromycin, the drugs have increasing frequencies after the 1990-2000 time period. In Figure 5, dirithromycin is only seen in the 1990s. It is a macrolide, like the standard macrolide erythromycin, as well as clarithromycin and azithromycin. Some studies showed that it may not offer any unique clinical advantage over clarithromycin or azithromycin (19).

Discussion

Infections with bacteria are important causes of morbidity and mortality worldwide. The control of bacterial disease requires a complex interplay of activities in the fields of public health, education, and medical science as well as in politics. There is a need for treatment, and the search for better drugs is a perpetual process. Advances in science and technology have opened up possibilities for new drugs.

In this study, the analysis of drugs for each bacterium highlights the frequency of usage decline and incline of liver drug therapies. Furthermore,

frequencies of liver drugs usage show the most popular drugs to treat a particular liver bacterium. According to these analyses, there are no big changes between time periods in the treatment of bacteria affecting liver. Despite the large global burden of bacterial diseases, there has been very little recent effort by the pharmaceutical industry to develop agents to treat human bacterial infections (20). The development of antibacterial drugs faces many challenges.

Antimicrobial resistance for bacteria threatens the management of infections, such as pneumonia, tuberculosis, malaria, and AIDS. During the past 10-15 years, antibiotic-resistant organisms have steadily increased, and now present a threat to disease management. In the past, resistance could have been handled by new drugs that are active against resistant microbes. However, the pharmaceutical industry has reduced its research efforts in agents against infections; genomics has not delivered the anticipated novel therapeutics; new regulatory requirements have increased costs; antibiotic use in common infections, e.g. bronchitis and sinusitis is questioned; and, compared with other drugs, return on investments is lower for microbials (21).

The protection of proprietary rights and the return of investments are also important issues for drug makers. With the long payback period associated with these indications, costs often are not recovered when a compound runs off patent and generic products may be introduced (22).

Regulatory requirements are another major concern that has a considerable impact on the length and costs of the drug development process and, hence, on the ultimate market price of the drug product. Paradoxically, increasingly demanding standards favour the larger wealthy companies, which are those least interested in tropical diseases. Nevertheless, dossiers do not always undergo the same level of review worldwide, sometimes because of limited health budgets, and sometimes owing to a misconception about the regulatory process (22).

Anti-infective agents differ from many other drugs in that treatment is normally given for a short time. By contrast with drugs used to treat hyperlipidemia, hypertension, and diabetes mellitus, for example, treatment of infections is rarely life long.

This short treatment period makes anti-infective drugs more susceptible to competition, the return per treatment course is limited, and the need for industry representatives increases. Marketing efforts generally continue during the entire life span of these drugs (21).

Considering all these concerns, from both clinicians' and patients' perspective, antibiotic resistance is an issue that continues to pose a significant threat. From the perspective of patients, headlines such as 'The revenge of the killer microbes' only add to their anxiety. As healthcare providers, they have a responsibility to acknowledge the issue of increasing resistance and to develop strategies for combating this continuing challenge to the management and treatment of infectious diseases (22).

In connection with this study, Medline abstracts were analysed. We noticed that although Medline includes biomedical articles dating back to 1950s, most articles in the 1950-1960 and 1960-1970 time periods do not contain abstracts. For example, some articles published in between 1960 and 1970 were retrieved from PubMed for analysis of *Mycobacterium tuberculosis*, but few drug names were extracted in these articles and the result does not provide sufficient information to make a comparison with other time periods. Therefore, this time period was removed from the time based analysis.

References

1. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association* 2008; 15: 87-98.
2. Wrongdiagnosis, <http://www.wrongdiagnosis.com>.
3. Zhou W, Smalheiser ND, Yu C. A tutorial on information retrieval: basic terms and concepts. *Journal of Biomedical Discovery and Collaboration* 2006; 1: 1-8.
4. Rebholz-Schuhmann D, Kirsch H, Gaudan S, Nenadic G, Arregui M. Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. In *Workshop on Multi-Dimensional Markup in NLP*, EAACL, Trento, Italy; 2006. .
5. Rebholz-Schuhmann D, Arregui M, Yepes AJJ, Kirsch H, Nenadic G. Automatic Text Analysis Based on Web Services. Handout for the ISMB 2007 Tutorial, ISMB, Vienna, 20.07.2007.
6. Drugbank, <http://www.drugbank.ca>.
7. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 2006; 1: 668-72.
8. Helms M, Ethelberg S, Molbak K, DT104 study group. International *Salmonella typhimurium* DT104 infections, 1992-2001. *Emerging Infectious Diseases* 2005; 11: 859-67.
9. Chen KL, Yeh CN, Lee HC. Analysis of the trend of drug resistance of *Salmonella typhimurium* in Taiwan, 1991-2001. *Epidemiology Bulletin* 2003; 19: 291-307.

Conclusion

Biomedical literature provides valuable knowledge for clinical studies and research. Medical experts cannot read all the articles in a specific medical problem and discover hidden connections between entities. We considered the need of medical doctors working on bacterial diseases to extract relevant drug and bacteria information embedded in published reports. We introduced a time based analysis of knowledge processing in Medline articles for the treatment of bacteria affecting liver. Biomedical text mining techniques were used to get useful facts embedded in Medline articles. Time based analyses of bacteria affecting liver-drug obtained in this study show that biomedical text mining techniques play an important role in extracting useful information contained in published articles. Therefore, both clinicians and researchers can use our proposed time-based comparison to improve treatment efficacy. This study also reveals that the treatment of bacteria seems to be stable over the past 4 decades and there are many challenges in introducing new drugs. The methodology introduced in this paper presents a reference model to acquire time based medical knowledge from the literature.

Acknowledgements

We would like to thank Antonio Jose Jimeno Yepes, Dietrich Rebholz Schuhmann, and George J. Towfic for their help and contributions.

10. Threlfall EJ, M Day M, Pinna E de. Goodyear Assessment of factors contributing to changes in the incidence of antimicrobial drug resistance in *Salmonella enterica* serotypes *Enteritidis* and *Typhimurium* from humans in England and Wales in 2000, 2002 and 2004. *International Journal of Antimicrobial Agents* 2006; 28: 389-395.
11. Quinolones, <http://www.dsir.gov.in/reports/techreps/tsr114.pdf>.
12. Michel M, Gutmann L. Methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant enterococci: Therapeutic realities and possibilities. *Lancet* 1997; 349: 1901-1906.
13. Nordmann P, Naas T, Fortineau N, Poirel L. Superbugs in the coming new decade; multidrug resistance and prospects for treatment of *Staphylococcus aureus*, *Enterococcus* spp. and *Pseudomonas aeruginosa* in 2010. *Current Opinion in Microbiology* 2007; 10: 436-440.
14. Danilchanka O, Mailaender C, Niederweis M. Identification of a novel multidrug efflux pump of *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy* 2008; 52: 2503-11.
15. Richeldi L, Covi M, Ferrara G, Franco F, Vailati P, Meschiari E et al. Clinical use of Levofloxacin in the long-term treatment of drug resistant tuberculosis. *Monaldi Arch Chest Dis* 2002; 57: 39-43.
16. Hidaka T. Current status and perspectives on the development of rifamycin derivative antibiotics. *Kekkaku* 1999; 74: 53-61.
17. Hsueh PR, Liu YC, So J, Liu CY, Yang PC, Luh KT. *Mycobacterium tuberculosis* in Taiwan. *J Infect* 2006; 52: 77-85.
18. Megraud F. *H pylori* antibiotic resistance: prevalence, importance, and advances in testing. *Gut* 2004; 53: 1374-84.
19. Wintermeyer SM, AbdelRahman SM, Nahata MC. Dirithromycin: A new macrolide. *Annals of Pharmacotherapy* 1996; 30: 1141-9.
20. White CA, Jr. Nitazoxanide: a new broad spectrum antiparasitic agent. *Expert Rev Anti Infect Ther* 2004; 2: 43-9.
21. Norrby SR, Nord CE, Finch R. Lack of development of new antimicrobial drugs: a potential serious threat to public health. *Lancet Infect Dis* 2005; 5: 115-9.
22. Trouiller P, Olliaro PL. Drug development output from 1975-1996: what proportion for tropical diseases. *International Journal of Infectious Diseases* 1998; 3: 61-63.