A comparison of structured data query methods versus natural language processing to identify metastatic melanoma cases from electronic health records

Jinghua He*

Merck & Co., Inc.,
Center for Observational and Real-World Evidence (CORE),
770 Sumneytown Pike, West Point, PA 19486, USA
Email: Jinghua_he@merck.com
*Corresponding author

Lawrence Mark

Indiana University School of Medicine,
Indianapolis, IN, USA
Email: lamark@iu.edu

Charity Hilton, Joel Martin and Jarod Baker

Regenstrief Institute,
Indianapolis, IN, USA
Email: Charity.Hilton@gtri.gatech.edu
Email: joel.martin443@gmail.com
Email: bakerjar@regenstrief.org

Jon Duke

College of Computing,
Georgia Institute of Technology,
USA
Email: Jon.Duke@gatech.edu

Siu L. Hui

Indiana University School of Medicine,
Indianapolis, IN, USA
and
Regenstrief Institute,
Indianapolis, IN, USA
Email: shui@iupui.edu

Xiaochun Li

Indiana University School of Medicine,
Indianapolis, IN, USA
Email: xiaochun@iu.edu

Paul Dexter

Indiana University School of Medicine,
Indianapolis, IN, USA
and
Regenstrief Institute,
Indianapolis, IN, USA
and
Eskenazi Health,
Indianapolis, IN, USA
Email: prdexter@regenstrief.org

Abstract: The relative efficacy of natural language processing (NLP) of text reports compared to structured data queries for identifying patients from electronic health records (EHRs) with metastatic cancer remains unclear. Such identification is critical for identifying and recruiting potential study candidates for cancer trials, particularly trials of cancer chemotherapy. For such purposes, we performed a direct comparison between NLP and structured data query methods for identifying patients with metastatic melanoma. Using EHR data from two large institutions, we found that NLP of text reports identified close to three times as many patients with metastatic melanoma compared to a structured data query algorithm (1,727 vs. 607 patients). Using an external tumour registry, we also found NLP had much higher sensitivity than structured query for identifying such patients (67% vs. 35%). Our results emphasise the importance of employing NLP criteria when identifying potential cancer study candidates with metastatic disease.

Keywords: efficacy; natural language processing; NLP; structured data queries; identification; identifying patients; electronic health records; EHRs.

Reference to this paper should be made as follows: He, J., Mark, L., Hilton, C., Martin, J., Baker, J., Duke, J., Hui, S.L., Li, X. and Dexter, P. (2019) 'A comparison of structured data query methods versus natural language processing to identify metastatic melanoma cases from electronic health records', *Int. J. Computational Medicine and Healthcare*, Vol. 1, No. 1, pp.101–111.

Biographical notes: Jinghua He obtained his PhD in Pharmaco Epidemiology from Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida. He also received his Master of Public Health degree from College of Public Health and Health Profession, University of Florida. He is currently a Principal Scientist at Center for Observation and Real-World Evidence, Merck Inc. & Co, Kenilworth, NJ, USA.

Lawrence Mark obtained his MD and PhD in Medical Biophysics from the Indiana University School of Medicine. He is actively involved in advancing medical research and medical education. He maintains a busy dermatological patient care load, specialising in melanoma and cutaneous lymphoma, at Indiana University Health in Indianapolis, IN.

Charity Hilton is a Research Scientist at the Georgia Tech Research Institute, leading the clinical natural language processing work at the Center for Health Analytics and Informatics. Her work has been focused on analytics within clinical domain for the past eight years. Prior to Georgia Tech, she served as a Lead Engineer of the Regenstrief Natural Language Processing Core. She is currently leading clinical NLP development on the open source platform, ClarityNLP, at the Georgia Tech Research Institute. She is currently pursuing a Master's Degree in Computer Science at Georgia Tech.

Joel Martin is currently a Data Scientist at Soostone. He was a data analyst and team leader at Regenstrief Institute, Indianapolis, IN, USA from January 2012 to August 2017.

Jarod Baker obtained his MBA in Healthcare Management from Indiana Wesleyan University. He currently serves as the Program Manager of the Merck-Regenstrief Institute collaborative partnership which is currently in its eighth year. The partnership seeks to improve the health of patients through data analytics, health care innovation, education and research.

Jon Duke is the Director of the Center for Health Analytics and Informatics at the Georgia Tech Research Institute and Principal Research Scientist in the Georgia Tech College of Computing. He has led over $25 million in funded research for industry, government, and foundation partners. His research focuses on advancing techniques for integrating, analysing, and communicating complex health data with applications spanning research, quality, public health, and clinical domains.

Siu L. Hui received her PhD in biostatistics from Yale University. She has been performing research in various areas of biomedical and health services research, a substantial amount of which has been based on electronic health records.

Xiaochun Li obtained her PhD in Statistics from the University of British Columbia. She has been specialising in electronic medical records linkage and databases, pharmaco-epidemiology studies, high-dimensional data, causal inference and analysis of observational studies, machine learning and statistical computing. In addition, she is well-versed in clinical trials with ten years of experience from working for Eli Lilly and Dana Farber Cancer Institute in clinical trials of Phases I-III.

Paul Dexter has been a Research Scientist at Regenstrief Institute for more than 20 years with a focus on adapting Regenstrief's information systems for both clinical and research purposes and has conducted multiple trials related to computerise clinical reminder systems. He served as the Chief Medical Information Officer at Wishard/Eskenazi hospital for 12 years. He helped implement a robust research IT infrastructure that includes tools related to decision support, natural language processing, and the efficient performance of identified and de-identified data queries. He currently serves as the Co-Chair of NHGRI's Implementing Genomics in Practice (IGNITE) consortia.

This paper is a revised and expanded version of a paper entitled 'A comparison of structured data query methods versus natural language processing to identify metastatic melanoma cases from electronic health' presented at International Society for Pharmacoepidemiology 32nd International Conference, Dublin, Ireland, 25–28 August 2016.

1 Introduction

Clinical trials are essential to determining the effectiveness of new cancer treatments, but less than 5% of adults with cancer enrol in such trials (Kehl et al., 2014). Others have demonstrated that clinical trial alert systems can efficiently increase clinical trial recruitment rates, but these systems rely on electronic health records (EHRs) as a means of automatically identifying study candidates who meet defined eligibility criteria (Embi et al., 2005; Thadani et al., 2009; Cuggia et al., 2011). With a goal of enhancing cancer trial enrolment rates, we compared the relative efficacy of natural language processing (NLP) of EHR text reports to EHR structured data queries for identifying patients with metastatic melanoma.

For many cancer trials, the presence or absence of metastatic disease is a critical distinction. Metastatic disease is a very important indicator of patient prognosis, as well as frequently serves as the primary inclusion criterion for trials of cancer chemotherapy. A small number of studies have examined the ability of NLP of text reports to identify such metastatic disease (Coden et al., 2009; Carrell et al., 2014; Nguyen et al., 2010; Spasić et al., 2014), whereas others have examined the ability of using ICD9 codes to identify metastatic disease (Chawla et al., 2014; Eichler and Lamont, 2009; Nordstrom et al., 2012; Quan et al., 2008).

Yet, the relative efficacy of NLP compared to EHR structured data queries (including ICD9 codes) for identifying metastatic disease remains unclear as a result of the small number of studies, inconsistent results among studies, and few direct comparisons. We describe below the results of such a direct comparison between NLP and structured data query methods for identifying patients with metastatic melanoma.

2 Materials and methods

2.1 Study design

We conducted a retrospective study to identify patients with metastatic melanoma from EHRs using two distinct approaches:

1 based on structured data queries

2 based on NLP algorithms applied to text reports.

The Indiana University Institutional Review Board approved this study.

2.2 Data sources

- *Identifying eligible patients from EHR data:* We used both structured and text report data from the Indiana Network for Patient Care (INPC) to identify patients with metastatic melanoma. The INPC, one of the nation’s oldest and largest health information exchanges, integrates patient data across participating institutions using global patient identifiers. Data available in the INPC include diagnoses, procedures, pharmacy claims, orders, laboratory results, and text reports such as discharge summaries, radiology reports, and pathology reports. We chose to limit our study to two of the oldest and largest clinical institutions in the INPC because:

1. those institutions had long sent the richest structured data to the INPC as a result of heavy clinical use
2. the text reports from those same two institutions had previously been indexed for Regenstrief's NLP platform.

- *External gold standard used for validation:* The Indiana Tumor Registry served as the external 'gold standard' source of metastatic melanoma cases. We used this external registry to estimate the sensitivities of the two algorithms under study. We included all cases of metastatic melanoma in the registry that could be matched to INPC global identifiers for patients with at least one clinical visit to one of the study institutions within one year before or after the diagnosis date in the registry.

2.3 Inclusion and exclusion criteria for the structured data query and NLP algorithms

The target population included all patients of age 21 years or older who had any clinical records in the INPC at the two study institutions between 1 January 2005 and 31 December 2013. Patients identified as prisoners were excluded.

- *NLP:* The NLP algorithm searched INPC text reports from the two institutions during the study years. Patients were judged to have NLP evidence of metastatic melanoma if they had text reports with the string 'melanoma' and derivatives of the word 'metastasis' (e.g., mets, metastases, and metastatic) within five-word phrases – with the following exception: we did not categorise five-word phrases as 'positives' when they met either of the following criteria:

  1. they were categorised as 'negated' or 'possible' using NegEx-derived software
  2. they were categorised as being in the family history section or otherwise associated with other family members or friends per Regenstrief's own family history annotator.

  For purposes of clarity, we considered patients who had both positive NLP evidence of metastatic melanoma in at least one text report, as well as negated instances (or in the family history) in either the same or other text reports, as affirmed or overall 'positive' cases of metastatic melanoma. In addition, we note that we did not use chemotherapy medication data as NLP criteria to identify patients.

- *Structured data query:* Eligible patients were included from the target population if they had at least one ICD-9 code (172.*) or local code consistent with melanoma, as well as a concurrent or subsequent ICD9 code (ICD9 198.* or 199.*) or local code consistent with metastasis. Such local codes were commonly derived from physician-maintained patient problem lists. Patients were excluded if they had cancer codes inconsistent with melanoma falling in time between the melanoma code and the metastasis code. We also included patients who had structured data (orders or claims data) indicating they had received a medication specific for treating metastatic melanomas (ipilimumab, vemurafenib, or trametinib) or had a BRAF test.

2.4 Validation through chart reviews

Chart reviews to establish the 'gold standards' used for estimating positive predictive values (PPVs) were conducted for all cases identified by the structured data query (which by necessity included all cases in the overlapping zone), as well as a 50% random sample of those identified by NLP alone. Charts reviewed in each zone were treated as a random sample of all identified charts within each zone. All of the chart reviews and consensus discussions discussed below were conducted in blinded fashion, without knowledge of the zone from which a particular chart was chosen.

We used two chart reviewers with clinical backgrounds (one a nurse practitioner, the other an overseas-trained physician) to review separate sets of sampled charts, with the initial charts reviewed by both until they consistently arrived at identical conclusions, as well as had the opportunity on several occasions to clarify details of the review process with the investigators using real patient examples. They were instructed to classify each chart as 'definite positive', 'definite negative', or 'unsure' for the occurrence of metastatic melanoma.

The chart reviewers were further instructed to classify the 'unsures' as 'unsure but possible' versus 'clinically no'. The 'unsure but possible' typically had too little reviewable information to make a definitive judgement call. The classic case of the 'clinically no' had many occurrences referring to an unrelated cancer (e.g., 'metastatic lung cancer'), but in a single text report, one sentence would refer to 'metastatic melanoma'. The chart reviewers, as well as Dr. Dexter, believed that these cases were almost undoubtedly a case of mistaken transcription and/or provider error.

After the independent reviews by each reviewer, a consensus review was conducted on all charts that were classified as 'unsure'. The chart reviewers were encouraged to consult with Dr. Dexter and among themselves for all open questions, with the former serving as the final arbiter on those very few cases (<5) for which the chart reviewers did not independently come to consensus.

2.5 Statistical methods

- *Estimation of PPVs:* We used the Venn diagram approach previously developed for a different phenotype in the INPC (Rosenman et al., 2014). Briefly, we drew a Venn diagram whose different zones contained cases of metastatic melanoma identified by NLP alone, by the structured data query alone, or by both. Samples were drawn from each zone for manual 'gold standard' chart reviews. The PPV (defined as the proportion among all identified cases that were ascertained to be true cases) of NLP and the structured data query were estimated with standard errors using weighted analyses based on the sampling weights.
- *Estimation of sensitivities:* The external 'gold standard' cases of metastatic melanoma came from the Indiana Tumor Registry, using registry patients who had records in the INPC indicating that they had received clinical care in at least one of the two study institutions within a year of the patients' tumour registry dates. All metastatic melanoma cases in the tumour registry were considered true cases. The sensitivity of each algorithm was defined as the proportion of external "gold

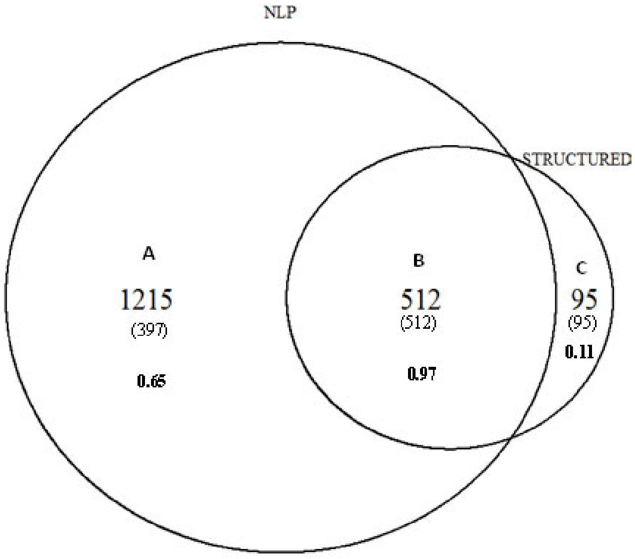
standard" patients that could be identified by the algorithm. Again, sampling weights were used in the estimation of sensitivities and their standard errors.

- *Comparisons between algorithms:* We used the 95% confidence intervals to compare the PPVs between the structured data query and NLP algorithms. PPVs were considered significantly different (nominally, that is, without multiple comparison consideration) if their 95% confidence intervals did not overlap.

3 Results

NLP identified a total of 1727 patients with metastatic melanoma (zones A and B in Figure 1). The structured data query identified a total of 607 patients with metastatic melanoma (zones B and C in Figure 1). A total of 512 patients were identified by both the structured data query and NLP (zone B in Figure 1).

Figure 1 A Venn diagram drawn to scale showing the number of unique patients identified as having metastatic melanoma, the number of charts reviewed for validation (in parentheses), and the PPV for the three zones



Notes: Zone A = patients identified only by NLP. Zone B = patients identified by both NLP and structured query. Zone C = patients identified only by structured query. For these purposes, we calculated PPVs only using 'definite yes' chart reviews as our estimate of true cases.

As summarised in Figure 1, the PPV for patients with metastatic melanoma identified by both NLP and structured data algorithms (PPV = 0.97, $p < 0.5$) was significantly higher than the PPV for patients identified only by NLP (PPV = 0.65) or only by the structured data query (PPV = 0.11).

The results from the 'gold standard' chart reviews are shown in Table 1.

Table 1 Number of charts reviewed and results of the chart reviews according to zone

Zone	*Number of charts reviewed*	*Definite yes (%)*	*Unsure, possible (%)*	*Unsure, clinically no (%)*	*Definite no (%)*
A	397	258 (65.0%)	33 (8.3%)	13 (3.3%)	93 (23.4%)
B	512	494 (96.5%)	4 (0.8%)	2 (0.4%)	12 (2.3%)
C	95	10 (10.5%)	3 (3.2%)	3 (3.2%)	79 (83.2%)

We found a total of 246 unique patients with metastatic melanoma in the Indiana Tumor Registry, with records in the INPC indicating that they received some clinical care in at least one of the study institutions within a year of the patients' tumour registry dates. These patients were used to estimate the sensitivities of the NLP and structured query algorithms, summarised in Table 2.

Table 2 Estimated sensitivities and PPVs for NLP, structured query, and either structured query or NLP

Zone	*Algorithm*	*Sensitivity (standard error)*	*PPV*	
			Definite yes	*Definite yes or unsure-possible*
A + B	NLP of text reports	0.67 (0.03)	0.74 (0.017)	0.80 (0.016)
B + C	Structured query	0.35 (0.03)	0.83 (0.015)	0.84 (0.015)
A + B + C	Structured query or NLP	0.67 (0.03)	0.71 (0.017)	0.77 (0.016)

Notes: For purposes of better understanding the potential effect of the 'unsure-possible' chart review category, we calculated PPVs both including and not including the 'unsure-possible' chart reviews in our estimates of true cases.

Using the registry-identified metastatic melanoma cases as the gold standard, we found the NLP algorithm to have a significantly higher sensitivity ($p < 0.05$) in detecting metastatic melanoma cases than the structured query algorithm, 67% vs. 35%.

If one uses only the 'definite yes' category as true cases, the PPV of the structured query algorithm was significantly higher ($p < 0.05$) than the NLP algorithm, 83% vs. 74%. However, if one uses both 'definite yes' and 'unsure but possible' as true cases, the difference between the PPVs of the structured query algorithm and the NLP algorithm (84% vs. 80%) does not reach statistical significance.

4 Discussion

Using data from two large institutions, we found that NLP of EHR text reports identified close to three times as many patients with metastatic melanoma as a structured data query that included ICD9 codes and local codes extracted from patient problem lists. Relatedly, using an external tumour registry, we found that NLP had much better sensitivity (67% compared to 35%) than a structured query. Our study highlights the importance of employing NLP techniques to identify patients who are potentially eligible for cancer trials, particularly trials of new chemotherapy agents for patients with metastatic disease.

The importance of using NLP techniques to identify potential cancer trial candidates using clinical trial alert systems is even greater when one considers that ICD9 codes are often assigned by an abstractor after the clinician has made therapeutic decisions. By using NLP as the clinician enters his or her initial note, one could alert the provider to an active cancer trial at the time that chemotherapy (for example) is being considered. Such clinical trial alert systems coupled with real-time NLP of clinician notes would represent a potentially powerful opportunity to improve on the less than 5% of adults with cancer who currently enrol in trials.

Our findings benefit from a study design that directly compared NLP to a structured data query, as well as being focused on a single form of cancer. Our study adds to a medical literature from which one could draw reasonably different conclusions about the relative efficacy of NLP to structured data queries to identify metastatic disease. For example, using ICD9 codes, Chawla et al. found sensitivities ranging from 43% for metastatic lung cancer to 73% for metastatic colorectal cancer, while Eichler found sensitivities of 97–100% for brain metastases. Similarly reflecting a wide range of results and possible conclusions, using NLP, Coden et al. (2009) achieved 32% sensitivity for metastatic tumour, while Carrell et al. (2014) reported a sensitivity of 92% for breast cancer recurrence that included metastatic disease.

For the study recruitment of most trials, one can accept a modest rate of error in identifying potential study candidates. That is, for most studies, one prefers to identify more potential study candidates (high sensitivity) even if such identification is occasionally wrong. This approach typically works owing to at least one subsequent screening step by either research assistants or clinicians to verify the study candidate's eligibility. For purposes of maximising the number of identified study candidates, we found NLP as the sole criteria to be the optimal approach, with a very reasonable PPV of 0.74–0.80 (depending on chart review criteria used). Nonetheless, we also found that the error rate could be minimised greatly by requiring both NLP and structured query evidence (PPV = 0.97, corresponding to the 'overlapping' zone B of Figure 1).

There are a few limitations to our study. While we did not use chemotherapy data in our NLP algorithm, we did include it in our structured queries. Because chemotherapy commonly exists in specialised oncologic information systems, it seems likely we had incomplete data (only four cases were found solely as a result of chemotherapy criteria in the structured query). It's certainly conceivable that we would have found more cases by structured query if we'd had more complete data, but alternatively, prior chemotherapy treatment is unlikely to be useful criteria if one's goal is to identify study candidates for new cancer chemotherapy trials.

We also conducted our trial using EHR data from two participating institutions, and only for metastatic melanoma. It would be useful to verify our findings at other external institutions and for other types of metastatic cancer. Finally, we also did not formally determine the inter-rater reliability of the chart reviewers, but rather worked to facilitate consensus among both the two chart reviewers and Dr. Dexter. By blinding all reviewers with respect to the zone that a particular chart represented, systematic biases are unlikely. In addition, our consensus process and multiple discussions that were centred on actual patients' charts were designed to ensure the highest likelihood that the final decision as to whether a patient had metastatic melanoma was correct.

We believe that the relative superiority of NLP to structured queries for finding metastatic melanoma cases is more important than the absolute sensitivities (67% vs. 35%). The Indiana Tumor Registry was an important external source of metastatic melanoma cases against which we could determine the relative sensitivities of the two approaches. However, we had no ready means of determining whether a large or only a small portion of a particular patient's medical care had been obtained at one of the study institutions. For patients who receive the bulk of their cancer care at a particular institution, it seems likely that the absolute sensitivities of both NLP and structured queries using that institution's EHR data would only be higher – while the relative superiority of the NLP approach seems likely to be maintained.

In summary, we found that NLP algorithms are superior to structured query algorithms for identifying patients with metastatic melanoma. Particularly given widespread adoption of EHRs, we believe that real-time NLP integrated with clinical trial alert systems represents a very promising method of increasing cancer trial enrolment.

Acknowledgements

The authors wish to thank and acknowledge Layla Baker and Karen Schmidt for their rigorous chart review efforts, which were critical for this study's success and also to David Haggstrom for his invaluable contributions to the current article.

References

Carrell, D.S., Halgrim, S., Tran, D., Buist, D.S., Chubak, J., Chapman, W.W. and Savova, G. (2014) 'Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence', *American Journal of Epidemiology*, Vol. 179, No. 6, pp.749–758.

Chawla, N., Yabroff, K.R., Mariotto, A., Mcneel, T.S., Schrag, D. and Warren, J.L. (2014) 'Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage', *Annals of Epidemiology*, Vol. 24, No. 9, pp.666–672.

Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K. and Groen, P.C. (2009) 'Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model', *Journal of Biomedical Informatics*, Vol. 42, No. 5, pp.937–949.

Cuggia, M., Besana, P. and Glasspool, D. (2011) 'Comparing semi-automatic systems for recruitment of patients to clinical trials', *International Journal of Medical Informatics*, Vol. 80, No. 6, pp.371–388.

Eichler, A.F. and Lamont, E.B. (2009) 'Utility of administrative claims data for the study of brain metastases: a validation study', *Journal of Neuro-Oncology*, Vol. 95, No. 3, pp.427–431.

Embi, P.J., Jain, A., Clark, J., Bizjack, S., Hornung, R. and Harris, C.M. (2005) 'Effect of a clinical trial alert system on physician participation in trial recruitment', *Archives of Internal Medicine*, Vol. 165, No. 19, pp.2272–2277.

Kehl, K.L., Arora, N.K., Schrag, D., Ayanian, J.Z., Clauser, S.B., Klabunde, C.N. and Keating, N.L. (2014) 'Discussions about clinical trials among patients with newly diagnosed lung and colorectal cancer', *Journal of the National Cancer Institute*, Vol. 106, No. 10, p.dju216.

Nguyen, A.N., Lawley, M.J., Hansen, D.P., Bowman, R.V., Clarke, B.E., Duhig, E.E. and Colquist, S. (2010) 'Symbolic rule-based classification of lung cancer stages from free-text pathology reports', *Journal of the American Medical Informatics Association*, Vol. 17, No. 4, pp.440–445.

Nordstrom, B.L., Whyte, J.L., Stolar, M., Mercaldi, C. and Kallich, J.D. (2012) 'Identification of metastatic cancer in claims data', *Pharmacoepidemiol Drug Saf.*, May, Vol. 21, No. 2, pp.21–8, doi: 10.1002/pds.3247.

Quan, H., Li, B., Saunders, L.D., Parsons, G.A., Nilsson, C.I., Alibhai, A. and Ghali, W.A. (2008) 'Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database', *Health Services Research*, Vol. 43, No. 4, pp.1424–1441.

Rosenman, M., He, J., Martin, J., Nutakki, K., Eckert, G., Lane, K. and Hui, S.L. (2014) 'Database queries for hospitalizations for acute congestive heart failure: flexible methods and validation based on set theory', *Journal of the American Medical Informatics Association*, Vol. 21, No. 2, pp.345–352.

Spasić, I., Livsey, J., Keane, J.A. and Nenadić, G. (2014) 'Text mining of cancer-related information: review of current status and future directions', *International Journal of Medical Informatics*, Vol. 83, No. 9, pp.605–623.

Thadani, S.R., Weng, C., Bigger, J.T., Ennever, J.F. and Wajngurt, D. (2009) 'Electronic screening improves efficiency in clinical trial recruitment', *Journal of the American Medical Informatics Association*, Vol. 16, No. 6, pp.869–873.