OPEN ACCESS

The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them

Mehmet Kayaalp, Allen C Browne, Fiona M Callaghan, Zeyno A Dodd, Guy Divita, Selcuk Ozturk, Clement J McDonald

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001689).

National Institutes of Health, National Library of Medicine, Lister Hill National Center for Biomedical Communications, Bethesda, Maryland, USA

Correspondence to
Dr Clement J McDonald, National Institutes of Health /National Library of Medicine, Lister Hill National Center for Biomedical Communications, 8600 Rockville Pike, Bldg. 38A/Room 7N707, Bethesda, MD 20894, USA; ClemMcDonald@mail.nih.gov

Received 31 January 2013
Revised 8 August 2013
Accepted 11 August 2013
Published Online First 11 September 2013

To cite: Kayaalp M, Browne AC, Callaghan FM, *et al*. *J Am Med Inform Assoc* 2014;**21**:423–431.

ABSTRACT

Objective To understand the factors that influence success in scrubbing personal names from narrative text.

Materials and methods We developed a scrubber, the NLM Name Scrubber (NLM-NS), to redact personal names from narrative clinical reports, hand tagged words in a set of gold standard narrative reports as personal names or not, and measured the scrubbing success of NLM-NS and that of four other scrubbing/name recognition tools (MIST, MITdeid, LingPipe, and ANNIE/GATE) against the gold standard reports. We ran three comparisons which used increasingly larger name lists.

Results The test reports contained more than 1 million words, of which 2388 were patient and 20 160 were provider name tokens. NLM-NS failed to scrub only 2 of the 2388 instances of patient name tokens. Its sensitivity was 0.999 on both patient and provider name tokens and missed fewer instances of patient name tokens in all comparisons with other scrubbers. MIST produced the best all token specificity and F-measure for name instances in our most relevant study (study 2), with values of 0.997 and 0.938, respectively. In that same comparison, NLM-NS was second best, with values of 0.986 and 0.748, respectively, and MITdeid was a close third, with values of 0.985 and 0.796 respectively. With the addition of the Clinical Center name list to their native name lists, Ling Pipe, MITdeid, MIST, and ANNIE/GATE all improved substantially. MITdeid and Ling Pipe gained the most—reaching patient name sensitivity of 0.995 (F-measure=0.705) and 0.989 (F-measure=0.386), respectively.

Discussion The privacy risk due to two name tokens missed by NLM-NS was statistically negligible, since neither individual could be distinguished among more than 150 000 people listed in the US Social Security Registry.

Conclusions The nature and size of name lists have substantial influences on scrubbing success. The use of very large name lists with frequency statistics accounts for much of NLM-NS scrubbing success.

INTRODUCTION

The personal name is one of the 18 identifiers defined by the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) that must be removed to de-identify the patient.[1] It is the most direct and natural way to identify a person. Personal name recognition has been a research interest for computational linguistics long before the medical community became interested in de-identification.[2]

De-identification of patient records has importance to clinical research, epidemiology, and medical informatics research. De-identification of well structured databases is relatively easy: one just removes the columns that carry fields named in the HIPAA Privacy Rule. The challenge arises with narrative reports: dictated clinical reports, typed physician visit notes, nursing notes, and radiology and other diagnostic study reports, which may contain embedded identifiers anywhere in their content.

Narrative clinical reports are of special interest to researchers because they contain information on symptoms, findings, and life events that are not usually available in structured parts of electronic health records. De-identification of such reports enables studies on large numbers of patients and complete populations that would be impossible or prohibitively expensive if dependent on identified data. For example, Kohane[3] proposes the use of de-identified medical record data and genetic information from discarded blood samples to obtain clues about genetic causes of rare diseases at a cost that is orders of magnitude lower and many-fold faster than prospective Genome Wide Association Studies (GWAS) studies, and Deleger *et al*[4] explain other research advantages of de-identified medical data.

Many researchers have developed and studied the effectiveness of de-identification tools using natural language processing (NLP) and statistical machine learning techniques. Mesytre[5] provides an excellent review. NLM has a long history of work with NLP tools[6–8] and has developed a de-identification tool called the NLM Name Scrubber (NLM-NS), based in part on NLM's existing NLP tools. In this report, we restrict our attention to the challenges of redacting personal names from narratives and we compare NLM-NS to four other name scrubbing/recognition programs.

BACKGROUND AND METHODS

Personal name scrubbing

Clinical reports may contain names of many kinds of persons including care providers and institutional staff. The latter are not the subjects of the health information,[9] so their names do not represent personal protected health information (PHI) per se. However, distinguishing between the names of patients and providers can be difficult, and provider names provide grist for testing name redaction methods; so NLM-NS, like most de-identification programs, attempts to remove all personal names. For simplicity of discourse, we call names of patients, their relatives, and household contacts, 'patient names' and names of hospital staff (eg, clinicians, physicians, nurses, and transcriptionists), 'provider names'.

Creating marked up reports for training and testing scrubbers

We derived reports for training and testing from a large set of HL7 V.2.3 observation messages[10] containing narrative clinical reports which we obtained from the NIH Clinical Center under an exemption from the NIH Office of Human Research Subjects Protection. HL7 messages are made up of structured segments analogous to records in a database. Each observation message begins with a number of header segments, for example, PID and PV1,[11] that precede any narrative report carried by that message. These header segment fields carry both patient and provider names and identifiers, which can be used to find and remove any such content from the associated narrative report.

HL7 messages are especially good targets for scrubbing systems because the majority of narrative clinical reports are delivered to electronic medical records (EMRs) in HL7 messages, and HL7 messaging is ubiquitous in large healthcare systems.[12–13] For example, all of the 120 hospitals interfaced to the Indiana Health Information Exchange (IHIE) (90% of the hospitals in Indiana) deliver HL7 messages to IHIE using HL7 messages (Shaun Grannis, Regenstrief Institute, written personal communication, May 31, 2013).

Test and training sets

We extracted our test and training sets of narrative reports from a 2-year set of 1.8 million HL7 observation messages provided by the NIH Clinical Center. We created a Clinical Center provider name list by extracting the names from the structured provider fields in the headers of these same messages. We created a Clinical Center patient name list by extracting the patient names from the HL7 message header segment (PID) linked to the reports of the study patients. What we call the Clinical Center name list is the union of those two.

When creating our test and training sets, we excluded laboratory test reports, which accounted for 86% of the HL7 messages, because they carried miniscule amounts of narrative that could carry personal names. We took our test and training sets from the subset of the 239 257 narrative reports contained in HL7 messages – dictated clinical notes and radiology reports – about 38 394 distinct patients. With software tools and manual review, a linguist and a nurse tagged every identifier (including names, IDs, etc.) in the reports used in our test and training sets. We tagged them using NLM's Visual Tagging Tool (VTT), an open source application written in JAVA that is freely available at http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/web/index.html. See figure 1 for the VTT display of a tagged fictitious report.

We created our set of test reports for this study by first taking a random sample of 1636 patients among the 37 370 patients with narrative reports who were not included in the training set. Then we took the last (in temporal order) of each report type for each patient in our sample. For clinician dictated reports, the report label (eg, discharge summary, operative note) defined the report type. For diagnostic studies, the individual study code (eg, chest x-ray) defined the report type. This process yielded a set of 3093 test reports. To exercise and improve the NLM-NS as we developed it, we used a set of 1140 annotated reports taken from narrative reports within the NIH Clinical Center HL7 messages. We call these our training set, which we also used to train LingPipe and MIST, machine learning systems. No patient or report from the training set was part of the test set described above.

We define a token as a string (of characters) separated by spaces or punctuation marks excepting the apostrophe which we treat as part of a name token. So, 'Mc Donald' represents two

Exam Date: 01/01/2012

REASON FOR STUDY:

Simone is a 93 year old woman with a history of metastatic breast carcinoma, currently on BMS-247550 infusion therapy. She refers new onset tingling in fingers of both hands and parethesias in dorsum of feet. Physical examiniation shows mild weakness (4+/5 according to the British Medical Research Council scale) of the first dorsal interossei (bilateral), left flexor digitorum profundus (4, 5), right abductor pollicis brevis and left tibialis anterior. She ahs gobal areflexial with loss of vibrations and pinprick in lower extremities (right side predominantly). This electrophysiological study is conducted to assess progress of sensory motor axonal polyneuropathy (documented last January). The patient has lymphedema in right lower and left upper extremities.

Findings:

1. Normal motor nerve conduction study of the right ulnar, and tibial nerves. The right median nerve still shows borderline conduction velocity without significant change in relation to the last study (conducted at Sinai Hospital 01/2004) rules out carpel tunnel syndrome). In comparison to the last study, the left peroneal nerve still shows reduced amplitudes of the motor response. All F-wave latnecies are normal.
2. Abnormal sensory nerve conduction study of the rht right median, ulnar, radial, and sural nerves with moderately reduced amplitudes of the sensory responses. Condution velocities are normal.

CONCLUSION:

This is an abnormal study. The electrophysiological findings are suggestive of a sensory motor (predominately sensory) axonal polyneuropathy.

Figure 1 Portion of a patient report after VTT tagging. Red signifies patient name, pink a numeric identifier, yellow a date, and green an age. This report includes only bogus PHI for demonstration purposes.

tokens whereas O'Leary and McDonald (no space) each represents one token. The tagging process labels both the full names (eg, John Quincy Adams) and stand-alone partial names (eg, 'Bobby' or 'Smith'), as name phrases. It also separately tags the individual components of these name phrases as tokens.

Scrubbing tools compared in this report

In this report, we compare a new NLM developed tool called NLM-NS and four existing open-source tools—MITdeid,[14] MIST,[15] LingPipe,[16] and ANNIE (based on GATE)[17]—that can recognize personal names in, and redact them from, narrative reports (see table 1).

Many scrubbers depend, in part, on predefined personal name lists to help scrub names within clinical documents.[5] A scrubber searches the document for strings that closely match any string in the name list and removes them. The name lists that came with three of these tools varied in size, derivation, structure, and the availability of name frequency statistics. Neither MIST nor LingPipe comes with an internal name list, but both can accept and use external name lists provided by the user.

The four existing systems are described in publications cited earlier.[14–17] NLM-NS used an algorithm that employs a combination of string matching, case checking, a weighing of the relative frequencies of a token in the NLM's mega name list versus its English word list, a list of prefixes (eg, Mr) and suffixes (eg, MD) and a Deterministic Finite State Automaton (DFSA)[18] to classify tokens in narrative text as names or non-names. Many scrubbers,[5] including MITdeid and ours, rely on regular expressions, which are mathematically equivalent to DFSAs[18] and can be compiled into DFSAs.

The NLM-NS algorithm first splits the narrative text portion of the report into sentences, and sentences into word, number, and punctuation tokens. It processes them in four steps. In the first step, it finds tokens following prefixes (eg, Mr, Dr, etc.) or preceding suffixes (eg, MD, PhD, Jr, etc.). These tokens are flagged as 'potential names'. In the second step, the algorithm looks up every token in the NLM mega name token list (3.8 million) derived from the person names in the Social Security Death Master File[19] and the Social Security registration file, and the author names from all Medline papers—and separately in its list of 2.5 million English words, which we derived from the words within the Wikipedia corpus and all words in Medline abstracts of the core clinical journals. See online supplementary appendix A for details regarding the development of these name and word lists and their issues. Tokens in all of these files are associated with a likelihood based on their prevalence within their corpus of origin. If a token's likelihood in NLM's mega name list was greater than its likelihoods in either of the two English word corpora *and* the word's initial letter was capitalized, it was marked as a potential name. Nobility tokens such as 'de', 'dos', and 'von' were classified as personal names, regardless of capitalization. If the token was capitalized and not found in any of the word or name lists, the algorithm also labeled it as personal name. In the third step the algorithm runs the tokens through a DFSA[18] to find multi-token patterns (see online supplementary appendix B). If a pattern includes one token marked as a potential name, and the average of the likelihood ratios of all tokens in the pattern (excluding prefixes, suffixes, and initials) is greater than 1, all of them are marked as potential names. The details are given as pseudocode in online supplementary appendix B. In the fourth step, the algorithm declares all potential name tokens that begin with upper case to be names. In study 2 and study 3, it also declares any tokens found in HL7 patient or provider name fields that are also present in the associated report to be names, regardless of case. In the last step, it finds in the report all unmarked tokens that were marked as names in other parts of the report and marks them as names. Throughout the development of NLM-NS, we focused on minimizing false negatives because that is what the HIPAA de-identification regulation demands.[20]

One of the goals of using such a large name list in the NLM scrubber was to see how successful a scrubber could be with a very large name list (eg, NLM's mega name list) without any customization by a local name list, though such customization may well be needed for optimal accuracy.

Detecting initials

Personal name initials occur as (1) stand-alone pairs, like 'JR', the moniker used by JR Ewing of the *Dallas* TV series, and triplets like 'JFK' (John Fitzgerald Kennedy), or (2) both as first or middle name initials (eg, E.E. Cummings or J Clement Stone) within full names. We tagged all patterns of initials in our gold standard report set as such. The DFSA was designed to find single initials and pairs of initials as part of a 'full' name, but we ignored them in our tally of scrubber performance, assuming failure to detect one initial (eg, 'F' in John F Kennedy) would not identify a patient if the rest of the name had been redacted.

Table 1 Scrubbing tools studied and information about their origins and availability

Name of scrubber	Source institution	Availability	Version tested
NLM-NS	National Library of Medicine, Lister Hill National Center for Biomedical Communications (Bethesda, MD)	Contact first author, Mehmet Kayaalp (mkayaalp@mail.nih.gov).	V1
MITdeid[14]	MIT (Cambridge, MA)	http://www.physionet.org/physiotools/deid/#software	*Tested version:* MITdeid V.1.1
MIST[15]	Mitre Corporation (Bedford, MA)	http://mist-deid.sourceforge.net/	*Tested version:* Mist V.1.2 (NLM Build). We later re-ran the study using MIST V.1.3.1, and got identical results to our run with MIST V.1.2.
LingPipe[16]	Alias-i, Inc. (Brooklyn, NY)	http://alias-i.com/lingpipe/web/download.html	*Tested version:* LingPipe V.4.1.0
ANNIE/ GATE[17]	University of Sheffield, Department of Computer Science (Sheffield, UK)	http://gate.ac.uk/download/ More information about ANNIE, which is distributed with GATE, is available here: http://gate.ac.uk/sale/tao/splitch6.html#chap:annie.	*Tested version* GATE V.6.0. (To confirm the validity of the reported studies, we later re-ran study 1 using GATE V.7.0 and obtained identical results for the patient names recognition, but slightly degraded results for provider name recognition, compared to GATE V.6.0)

We also ignored 'stand alone' initials like 'LBJ' (Lyndon Baines Johnson) in all of the scrubber comparisons, for simplicity's sake. However, we did assess and report the detection of both kinds of initials separately.

Testing of the scrubbers

We tested the scrubbers in three studies. In the first, all of the systems used only their out-of-the-box native name list. NLM's mega name list was the native name list for NLM-NS. In the second study, we added the Clinical Center provider list to the native name list of all the systems. We gave the two systems, NLM-NS and MITdeid, which could process patient names specific to each report, access to the tightly-linked patient names in the report's HL7 PID header. The version of the MITdeid scrubber that we tested accommodates only one directly tied full *patient* name per report, whereas NLM-NS could take multiple names per report. To put the systems that could not redact specific names by report on an even footing with NLM-NS and MITdeid, we added all of the Clinical Center patient name token list (all patient name tokens in the HL7 header segments of study patients) to their native name list.

We ran a third study that added the NLM's mega name list to the name lists included in the second study and tested the three systems that could accommodate that large list.

In each study, we compared each scrubber's classification of each test report token as a personal name or not against our hand-annotated gold standard classification of that token. In the analysis, we did not count prefixes (eg, Mr, Mrs, Dr) or suffixes (eg, MD, PhD) as name tokens, though most algorithms used them to help decide what to scrub.

Data analysis

We scored the success of each scrubber as true and false positives (ie, TP, FP) and true and false negatives (ie, TN, FN) in the usual two-way table based on the total number of tokens (excluding punctuation and single letters). In post-study analysis we separately examined the frequency of full names and initials in the test reports and assessed the degree to which scrubbers fail to redact them. We also report the number, percentage, and type of names within the test set of reports that could potentially be detected by simple string matching with the names in each of the name lists employed in these studies. During the analysis, our primary focus was on sensitivity (minimization of false negatives) as has been the focus of most de-identification systems.[5 15 21] We also report specificity, and for the sake of completeness, F-measures.[22]

To compare the overall test of the difference among the sensitivities or specificities of the scrubbers in the various test, we used a K-sample test of proportion (Pearson's χ^2 d statistic)[23] for all six cases (patient token, provider token, all token, patient unique token, provider unique token, all unique token). We then tested whether the NLM-NS method was better than each of the other methods 'head-to-head' in a series of EXACT two-sample tests of proportion. As all these tests were specified a priori, there was no adjustment for multiple testing. The testing was performed using the BINOM.TEST package in the statistical software R.[24] All CIs and hypothesis tests used a significance level of 5%.

RESULTS

Test set characteristics

The test set included 3093 clinical reports about 1636 distinct patients, all of whose full names, as represented in the HL7 header, were distinct. The number of reports per patient within the test set ranged from 1 to 20 (mean 1.9). Diagnostic services were the source of 65% of the test reports, with an average size of 154 words, and most (96%) of these were radiology reports. The remaining 35% of the clinical reports were 19 different types of provider dictated reports. These reports had an average size of 759 words; see table C1 in online supplementary appendix C for the distribution of report types.

These 3093 test reports contained 1.1 million token instances, of which 22 548 (2%) were personal name (patient or provider name) tokens. Of the personal name tokens, 2017 were unique, and of the non-name tokens, 38 922 were unique. Of the unique personal name tokens, 141 (7%) overlapped with non-name tokens. The majority, 20 160 (89%), of name token instances came from provider, and a minority, 2388 (11%), from patient names. All of the patient name tokens began with an upper case letter. Forty-three percent of them were preceded by a title (Mr, Ms); none included a suffix. And 98% of the unique patient name tokens had higher likelihoods in the NLM mega name list than in either of the two English word lists.

The full set of reports contained 2012 instances of patient *name phrases* and 751 unique such phrases, counting multi-word, for example, 'Robert Johnson', and single-word, for example, 'Bobby', phrases presented in different parts of a report as separate name phrases. The number of patient name phrases in the 638 test reports that carry any patient name phrase ranged from 1 to 49 (a mean of 3.2). The report with the maximum number of name phrases appeared to include every family member's name. Most patient name phrases (83%) consisted of *single* name tokens, either the last name (eg, Mr Smith) or the first name (eg, John). The test reports carried 2388 patient name *token instances* and 745 unique such tokens. The 638 reports that carried at least one patient name token had a mean of 3.7 (max. 51) instances of such tokens. Names of patient relatives represented a small proportion, 3.2%, of the 'patient name' token instances and 2.3% of the unique such tokens. The maximum number of tokens occurring within one name phrase was four.

Half of the provider generated reports carried patient name tokens (2.5 such token instances per report), but only 6% of diagnostic study reports carried such tokens (0.2 token instances per report). The name tokens carried by diagnostic study reports were concentrated in 173 DEXA and 5 Holter reports. Patient name tokens occurred in the body of less than 1% of all other diagnostic study reports.

A total of 10 154 instances of provider name *phrases* (a mean of 3.3 per report) and 6759 unique such phrases (a mean of 2.2 per report) appeared in the test reports. In stark contrast to the case for patient name phrases, 94.6%, the provider name phrases consisted of *multiple* tokens, for example, 'William Osler', not counting titles like Dr and MD. Unsurprisingly, provider names occurred within all provider dictated reports and all but three diagnostic study reports, averaging 6.5 provider name token instances and four unique such tokens per report. Provider name tokens came from mentions of clinicians, nurses, transcriptionists, diagnostic study interpreters, and other institutional personnel.

Initials

One hundred and seventy-two instances of stand-alone persons' initials, including 53 triplets like JFK, 118 doubles (eg, JR of JR Ewing), and one single ('Dr J') existed in the set of test reports. Of these, 34 were unique, all but one of the 172 were provider names, and all represented the person who dictated or transcribed the report. Only one of the stand-alone patient initials

was a double initial, which NLM-NS recognized as a name. We did not count detection success for initials in our tally of sensitivity because they were outside of our planned scope. However, each of the tested systems scrubbed some of providers' stand-alone initials. NLM-NS identified 48%, MIST 39%, LingPipe 27%, and MITdeid and ANNIE 1% of the unique initial instances. These failures are not crucial because provider names and their initials do not represent PHI.

We deliberately ignored single initials that were part of a name phrase in our study scope, but many of the systems including NLM-NS did detect them. Single initials occurred in patient name phrases 105 times. One pair of initials separated by a space and/or period that appeared twice accounted for a total of four single letter initials embedded in names. Single middle initials accounted for the rest. We did not include counts of one letter initials scrubbing success in our primary analysis but the NLM scrubber did detect them all.

Scrubbers' performance

In all of our analyses we report the scrubbing success based on token instances and on unique tokens. The measure by instances is the most optimistic because it exaggerates the effective sample size. The result, based on unique tokens, is the most conservative and ignores differences in context such as the presence of a title (Mr) that might correctly identify a token as a name when preceded by that title and miss it when not. LingPipe failed to scrub 22 instances of one nickname, Charlie. The first study tested scrubbing performance at the token level for all five systems using their out-of-the-box name lists. Results are given in table 2.

In this first analysis NLM-NS failed to detect only two patient name tokens and revealed no full patient names. Its two failures were a nickname of a spouse and a last name of another patient that contained typos at two different positions within the name. Out of the box (using their native name lists) the next two best scrubbers failed to identify 145 and 375 patient name tokens. For patient and provider name token instances and unique tokens, NLM-NS missed fewer name tokens than did any of the other scrubbers; the differences were significant for all comparisons ($p<0.05$) except for the comparison of its number of missed provider name token instances with MIST ($p=0.201$). For the specificity of all tokens, MITdeid had the best specificity for both token instances and unique tokens ($p<0.0001$) compared to each of the others systems.

In the second study we gave MITdeid and NLM-NS access to the patient names tightly linked to each report and added the Clinical Center provider token name list (a total of 4801 names) to their native name lists. ANNIE, LingPipe, and MIST could not process name lists linked tightly to specific reports. So we added the Clinical Center provider name list *and* all of the tightly linked patient name tokens (2677 of them) for a total of 6619 unique tokens added to their native name lists.

The performance of all of the non NLM-NS systems improved markedly in this second study (table 3) with the availability of names derived from the same institution as the test reports. For example, the number of false negatives for patient name tokens was reduced by 92% for MITdeid (from 145 to 11) and improved for LingPipe, MIST, and ANNIE by 93%, 78%, and 68%, respectively. A similar pattern was seen for unique patient names. The redaction of provider names by most systems generally improved even more dramatically. MIST was unique in showing a large improvement in specificity as its sensitivity improved. In study 2, the additional patient name list information had no effect on NLM-NS's patient name redaction because its mega name list was so inclusive; however, the additional provider names did improve its redaction of provider names by 45%.

The overall test of the difference among the sensitivities and specificities for all six cases (patient token, provider token, all token, patient unique token, provider unique token, all unique

Table 2 Study 1: experimental results at the token level tallying scrubbing success when scrubbers used their native name list

Token instances	Patient tokens		Provider tokens		All tokens (non-name tokens)		
Total N	2388		20 160		1 126 241 (1 103 693)		
	Sensitivity (95% CI)	FN	Sensitivity (95% CI)	FN	Specificity (95% CI)	F-measure	FP
NLM-NS	0.999 (0.997 to 1)	2	0.999 (0.999 to 1)	11	0.987 (0.987 to 0.987)	0.756	14 510
MITdeid	0.939 (0.929 to 0.948)	145	0.850 (0.845 to 0.855)	3027	0.998 (0.998 to 0.998)	0.871	2580
MIST	0.843 (0.828 to 0.857)	375	0.999 (0.998 to 0.999)	19	0.993 (0.993 to 0.993)	0.848	7573
LingPipe	0.829 (0.813 to 0.844)	409	0.978 (0.976 to 0.980)	443	0.954 (0.953 to 0.954)	0.456	50 905
ANNIE	0.825 (0.809 to 0.840)	417	0.741 (0.735 to 0.747)	5221	0.989 (0.989 to 0.989)	0.659	11 893

Unique tokens	Patient tokens		Provider tokens		All tokens (non-name tokens)		
Total N	745		1449		40 798 (38 922)*		
	Sensitivity (95% CI)	FN(FN+TP)	Sensitivity (95% CI)	FN(FN+TP)*	Specificity (95% CI)	F-measure	FP(FP+TN)*
NLM-NS	0.997 (0.989 to 1)	2 (745)	0.996 (0.991 to 0.998)	6 (1451)	0.942 (0.940 to 0.944)	0.634	2311 (39 909)
MITdeid	0.908 (0.885 to 0.928)	70 (765)	0.820 (0.800 to 0.839)	277 (1542)	0.985 (0.984 to 0.987)	0.796	573 (39 169)
MIST	0.898 (0.875 to 0.918)	82 (805)	0.988 (0.980 to 0.992)	18 (1460)	0.956 (0.954 to 0.958)	0.681	1765 (40 412)
LingPipe	0.752 (0.721 to 0.780)	211 (851)	0.877 (0.860 to 0.893)	198 (1614)	0.824 (0.821 to 0.828)	0.316	7801 (44 403)
ANNIE	0.756 (0.724 to 0.785)	188 (770)	0.764 (0.743 to 0.784)	385 (1633)	0.978 (0.977 to 0.980)	0.702	855 (39 368)

We report false negative results for patient names and provider names separately and false positive results for the patients and providers combined.
*For measures based on unique tokens, we had to use different FN+TP values per scrubber, rather than these totals, because when comparing uniques, a given token could be a FP in one context and TP in another, and the classification could change by scrubber.
FN, false negative; FP, false positive; TN, true negative; TP, true positive.

token) in study 2 was significant with p values all at <0.0001, indicating some differences among the scrubber methods. In this study, NLM-NS had the fewest false negatives in both name token instances and unique tokens of provider and patient names, and these differences were significant for comparisons with non-NLM systems for all patient and provider name token instances except for the comparison with MIST for provider instances (p=0.12). NLM-NS also had significantly fewer failures to scrub *unique* patient and providers tokens for all comparisons except for MITdeid (p=0.29) and LingPipe (p=0.68). In study 2, for all token instances, MIST had the best specificity (0.997) and best F-measure (0.938). For unique token instances in study 2, MITdeid had the best specificity (0.980) and best F-measure (0.830). Except for LingPipe, the specificities were all greater than 0.94. For privacy concerns, however, minimizing false negatives is more 'important' than false positives.[21]

The effect of the NLM mega name list on the scrubbing success of de-identification systems

In all of the comparisons NLM-NS employed its mega name list as its native name list. To determine whether this list would benefit the other scrubbers as well, we ran another test in which we added NLM's mega name list to the name lists used in study 2. MITdeid, MIST, and ANNIE could digest this large list. LingPipe could not, and therefore was excluded (see table 4).

The sensitivity of both MITdeid and ANNIE improved considerably with the addition of the mega name list. Indeed with this addition, MITdeid's false negative performance exactly equaled that of NLM-NS assessed on *unique* tokens, though the two systems failed on two different pairs of tokens. As implemented in this study, neither MITdeid nor ANNIE could take advantage of the name frequencies in the mega name list, and therefore they simply removed all tokens within the mega name list from the test reports—see table 4 for the detailed results.

Because of the enormous overlap of the rare names in the mega name list with ordinary words, these two systems removed from half to two-thirds of all of the tokens in these reports. MIST behaved differently from the aforementioned two. In study 3, MIST's sensitivity for patient names worsened and was unacceptable, but its specificity held up (with 99.4% for all token instances and 96.3% for all unique tokens), and was better than any other system. NLM-NS use of the frequency of tokens in its mega name list versus that frequency in the NLM corpus of English words protected it from such damaging effects on specificity, which some of the other systems suffered. MITdeid has the option for distinguishing popular, unambiguous names and ambiguous names by gender, but we did not take advantage of this feature in this study. MIST had worse sensitivity than NLM-NS for both patient and provider names, whether counted as unique tokens or token instances (all p<0.05).

Name list coverage of tokens in the test set

The success of name recognition systems depended in part on the nature and size of their name lists. The native name lists associated with three of the de-identification system carried 41% (ANNIE), 80% (MITdeid), or 99.5% (NLM's mega list) of the unique patient name tokens within the test reports. With one exception, their coverage of tokens in the test reports increased with the name list size—see table C2 in online supplementary appendix C. However, the Clinical Center name list, which carried only 6619 distinct name tokens, 2677 of which derived from study patient tightly linked names, included 96.9% of the unique patient name tokens within our test set because it was so specific to the population from which the test

Table 3 Study 2: results of comparison of personal name removal when the compared systems had access to the name list derived from the Clinical Center data and the names in the HL7 header segments directly linked to the report

Token instances	Patient name tokens		Provider name tokens		All tokens (non-name tokens)		
Total N	2388		20 160		1 126 241 (1 103 693)		
	Sensitivity (95% CI)	FN	Sensitivity (95% CI)	FN	Specificity (95% CI)	F-measure	FP
NLM-NS†	0.999 (0.997 to 1)	2	1 (0.999 to 1)	6	0.986 (0.986 to 0.986)	0.748	15 214
MITdeid◇	0.995 (0.992 to 0.998)	11	0.999 (0.999 to 0.999)	18	0.983 (0.983 to 0.983)	0.705	18 835
MIST‡	0.965 (0.957 to 0.972)	83	0.999 (0.999 to 1)	14	0.997 (0.997 to 0.997)	0.938	2876
LingPipe‡	0.989 (0.983 to 0.992)	27	0.999 (0.998 to 0.999)	30	0.935 (0.935 to 0.936)	0.386	71 619
ANNIE‡	0.944 (0.934 to 0.953)	133	0.938 (0.935 to 0.941)	1247	0.983 (0.983 to 0.983)	0.676	18 939

Unique tokens	Patient name tokens		Provider name tokens		All tokens (non-name tokens)		
Total N	745		1449		40 798 (38 922)*		
	Sensitivity (95% CI)	FN(FN+TP)*	Sensitivity (95% CI)	FN(FN+TP)*	Specificity (95% CI)	F-measure	FP (FP+TN)*
NLM-NS†	0.997 (0.989 to 1)	2 (745)	0.997 (0.992 to 0.999)	4 (1450)	0.940 (0.938 to 0.942)	0.626	2394 (39 962)
MITdeid◇	0.992 (0.982 to 0.997)	6 (746)	0.988 (0.980 to 0.992)	18 (1451)	0.980 (0.978 to 0.981)	0.830	792 (39 157)
MIST‡	0.959 (0.942 to 0.972)	31 (765)	0.990 (0.984 to 0.995)	14 (1457)	0.979 (0.978 to 0.981)	0.822	824 (39 628)
LingPipe‡	0.995 (0.985 to 0.998)	4 (747)	0.984 (0.975 to 0.989)	24 (1457)	0.817 (0.814 to 0.821)	0.330	8094 (44 269)
ANNIE‡	0.865 (0.838 to 0.888)	103 (762)	0.883 (0.866 to 0.898)	185 (1582)	0.972 (0.970 to 0.974)	0.732	1098 (39 419)

False negative rates reported separately for patients and providers, and false positive rates reported together.
*For measures based on unique tokens, we had to use different FN+TP values per scrubber, rather than these totals, because when comparing uniques, a given token could be a FP in one context and TP in another, and the classification could change by scrubber.
†Access by report to patient and provider names from the HL7 header.
◇Access by report to patient names from the HL7 header and all providers.
‡Access to all patient names in study and to provider names.
FN, false negative; FP, false positives; TN, true negative; TP, true positive.

Table 4 Study 3: token level performance of systems after adding NLM mega name list plus Clinical Center name list to the native name list of MITdeid, MIST, and ANNIE

Token instances	Patient tokens		Provider tokens		All tokens (non-name tokens)		
Total N	2388		20 160		1 126 241 (1 103 693)		
	Sensitivity (95% CI)	FN	Sensitivity (95% CI)	FN	Specificity (95% CI)	F-measure	FP
NLM-NS†	0.999 (0.997 to 1.000)	2	1.000 (0.999 to 1.000)	6	0.986 (0.986 to 0.986)	0.748	15 214
MITdeid◇	0.997 (0.994 to 0.999)	7	1.000 (1.000 to 1.000)	3	0.318 (0.317 to 0.319)	0.057	752 437
MIST‡	0.871 (0.857 to 0.884)	307	0.999 (0.999 to 0.999)	17	0.994 (0.994 to 0.994)	0.870	6 338
ANNIE‡	0.969 (0.962 to 0.976)	73	0.942 (0.939 to 0.946)	1162	0.498 (0.497 to 0.499)	0.071	553 854

Unique tokens	Patient tokens		Provider tokens		All tokens (non-name tokens)		
Total N	745		1449		40 798 (38 922)†		
	Sensitivity (95% CI)	FN (FN+TP)*	Sensitivity (95% CI)	FN (FN+TP)*	Specificity (95% CI)	F-measure	FP(FP+TN)*
NLM-NS†	0.997 (0.989 to 1.000)	2 (745)	0.997 (0.992 to 0.999)	4 (1450)	0.940 (0.938 to 0.942)	0.626	2394 (39 962)
MITdeid◇	0.997 (0.989 to 1.000)	2 (747)	0.998 (0.993 to 0.999)	3 (1450)	0.669 (0.665 to 0.674)	0.230	13 469 (40 727)
MIST‡	0.909 (0.887 to 0.928)	73 (806)	0.989 (0.982 to 0.994)	16 (1461)	0.963 (0.961 to 0.965)	0.717	1491 (40 178)
ANNIE‡	0.933 (0.912 to 0.949)	52 (774)	0.888 (0.872 to 0.903)	177 (1586)	0.792 (0.788 to 0.796)	0.308	8588 (41 302)

*For measures based on unique tokens, we had to use different FN+TP values per scrubber, rather than these totals, because when comparing uniques, a given token could be a FP in one context and TP in another, and the classification could change by scrubber.
†Access by report to patient and provider names from the HL7 header.
◇Access by report to patient names from the HL7 header and all providers.
‡Access to all patient names in study and to provider names.
FN, false negative; FP, false positives; TN, true negative; TP, true positive.

reports were taken. The Clinical Center name list included a smaller proportion, 87.9%, of the *provider name* tokens because HL7 message headers do not carry information about all of the providers mentioned in the reports.

The effect of tightly linked names

The HL7 header segments contain full patient and provider names that are tightly linked to its corresponding report. The name tokens in a report's HL7 header segments covered only 92.1% of the unique patient name tokens *within the report itself* and 83.4% of the corresponding provider name tokens. We reviewed all of the reports for which the tightly linked patient names in the HL7 header did not identify all of the patient name tokens in its report. We found 188 instances of such tokens in 56 reports or 7.8% of the 2388 patient name token instances. Of these, 41 (22%) were misspellings, 43 (23%) truncations, 47 (25%) patient nicknames, 52 (28%) family members, 3 (2%) maiden name of the patient, 1 (1%) was a wrong name, and 1 (1%) a plural form of the patient's family name. NLM-NS found all but two of these tokens.

Confidence in scrubbing success

NLM-NS failed to detect two patient name tokens in two different patients across 1636 unique patients and 3093 reports. By strict Safe Harbor rules, these represent two failures. However, the nickname that NLM-NS failed to scrub is a nickname for more than 200 000 people in the US according to US Social Security name data (frequencies of names of applicants for a US Social Security Number). Furthermore, the two typographical errors on the second 'failure' would include more than 150 000 individuals if expanded to all of the possible names within the spell-check distance between the mangled and the correct name based on the Social Security name data. Thus, statistically, neither of these strict 'failures' is a real failure to protect privacy.

The space in which scrubbing results are interpreted is complex and contains much interdependence: patients have multiple reports; reports may contain multiple names; names may consist of many tokens; the same tokens can appear as part of more than one name; and the recognition of tokens can depend on context. Consequently, many choices for the denominator exist, for example, number of reports, number of unique patients, number of token instances, and number of unique tokens. Therefore, we analyzed the data in terms of unique token and token instances as the worst and best case assumptions.

In this analysis, the raw failure rate of NLM-NS was one per 373 patient name tokens, one per 818 patients, and one per 1547 reports. However, even under the best case assumption of no strict PHI failures (considering the lower bound of the corresponding CI), we can only be sure we won't miss more than one personal name per 202 unique name tokens, per 444 (patients), and per 839 (reports), at the over 95% confidence level.

DISCUSSION

An operational scrubber would need to redact many kinds of identifiers and most published reports of de-identification systems included multiple types of PHI (eg, names, addresses, identification numbers). This report focused only on the scrubbing of names in order to better understand the factors for successful name scrubbing.

The NLM-NS removed more *patient* name token instances than the other scrubbers in all comparisons at a reasonable level of specificity. It found all but 2 of 2388 (0.08%) patient name token instances (one of which was a relative's nickname and the other was another patient's last name misspelled at two positions) which did not expose PHI under HIPAA's statistical rule.

When given access to a name list from the institutional source of the test reports (study 2), the scrubbing performance of the other systems improved substantially.

In this effort, our primary goal was the *removal* of personal names, which meant achieving a false negative rate as close to zero as possible, and we sacrificed specificity to achieve this end. In study 2, for all patient name instances, we ended with a specificity of 0.986, implying 1.3% of the words in the document would be incorrectly removed. We believe this was an acceptable price to pay for greater privacy protection; it is in the range of most published de-identification studies. Although NLM-NS always had the lowest absolute false negative rate for both patient and provider name tokens (whether counted as instances or unique tokens), NLM-NS was not significantly better than MITdeid or LingPipe for detecting *unique* patient names, and MITdeid and MIST had better specificity than NLM-NS. We wish to emphasize that the results we found for the non-NLM scrubbers should not be taken as the best they can do; the original developers will know more ways to tune them than we did.

Four factors influenced the NLM-NS's detection of personal names classification: (1) the likelihood ratio (between NLM's mega name list and its English word list); (2) the presence of an adjacent prefix or suffix; (3) for study 2, the occurrence of personal name tokens in the body of a report that were also within a name field of a header segment for that report; and (4) the DFSA. The influence of these factors depends on the order in which the program happens to address them. But considered independently, each of the first three factors are strong discriminants. All but 16 of the 745 unique patient name tokens (98%) would have been correctly classified as names by the likelihood ratio alone. A rule that checked for a preceding prefix and upper case would have found 42% of the personal names. The patient name strings in the HL7 header would have correctly identified 92% of the patient name tokens in the report. So there are many routes to 'name-hood'. The DFSA was a mop-up operation to detect words that would not be recognized as personal names except for their collocation with words that had already been classified as personal names.

We were pleased with the very low failure rate of NLM-NS on a much larger sample size of distinct patients (1636) than most previous studies, one of which had a much lower sensitivity for patient name failure rate[25] and one of which did not report the number of distinct patient name phrases or tokens.[26] Our study was very similar in size and scope to Deleger's study,[4] with almost identical numbers and varieties of reports, total numbers of tokens (both close to 1 million), and both included MIST in their comparisons, but NLM-NS was a more effective personal name scrubber. The test reports in this study were all produced by transcriptions, who paid careful attention to name capitalization, a feature on which our scrubber depended. So the results in this study will not apply to text with inaccurate capitalization as may occur in notes entered directly by care providers.

On the other hand, even with our sample size, we could not be certain at the lower 95% confidence limits that NLM-NS would miss fewer than one patient name token per 204 unique such tokens and per 833 unique patients within the test reports, even when assuming the two surface failures were not real failures. The nature of the binomial distribution means that sample reports from very large numbers of patients are required to prove the success of scrubbers at a high degree of statistical confidence. For example, a sample of reports carrying 7375 distinct patients would be needed to be sure that a scrubber with no failures would not miss more than one personal name token per 2000 patients at the 95% confidence level. Most studies of scrubbing success have included far smaller numbers. For example, the studies in the I2B2 challenge[27] included at most 75 distinct patients, which we calculated from 98.6% sensitivity Uzuner *et al* reported in table 11 of their data supplement,[27] when one patient was missed. Therefore, even if the best performing I2B2 system missed no patient names, it could have missed as many as one name per 21 patients at the lower 95% confidence limits.

The nature and content of name lists is obviously a very important factor in the success of de-identification tools. Neamatullah and colleagues[14] emphasized that the regular expressions and name patterns alone could detect >90% of the personal names. But string matching alone, with even modest sized name lists, such as the Clinical Center name list, can achieve even better, 96.9%, performance levels— see table C2 in online supplementary appendix C. NLM-NS's mega list, where each name was linked to a frequency, was a major factor in NLM-NS's success. Such large, comprehensive name lists provide good scrubbing without the need for institution specific name lists and might enable scrubbing as a web service. Interestingly, even this huge list did not include 0.5% and 1.1% of the unique patient and provider name tokens, respectively. The success of NLM-NS depended largely on its comparison of token likelihood between its mega name list and its English word list. However, we believe this part of our algorithm still has room for improvement. The total Wikipedia corpus was so rich with personal names that the nickname we missed was more prevalent in the Wikipedia English word list than in NLM's mega name list.

The prevalence of provider names was much higher than that of patient names in this study, as has been the case in most studies of scrubbers. The patterns of these two kinds of names within clinical reports are quite different. Provider names are usually full names associated with prefixes and titles (eg, MD), often preceded with labels, for example, 'signing physician'. In theory, those labels and cues should make provider names easier to detect than patient names, which are most often single token names (first or last names), and usually appear without distinguishing titles or labels. So, we should not assume that provider name scrubbing success predicts patient name scrubbing success, though many studies have not distinguished their scrubbing success across these two categories.

Our algorithm did not convert patient first names found in HL7 name fields to their associated nicknames to help find patient nicknames embedded in our test reports, but doing so would not have improved NLM-NS's success because neither of NLM-NS's failures were *patient* nicknames. However, the addition of nicknames for names in local name lists would generally be expected to help scrubbing success.

As tested, none of the systems were good at finding stand-alone initials. However, all but one of these patterns of initials were those of providers, and not PHI. Whether the inclusion of patient initials in a patient report could identify the patient is an open question, but there is probably an easy method for removing them when text reports come from HL7 messages: the scrubber could generate all 2–3 letter abbreviations for the tightly linked patient names in the header segment and then strip such strings from target reports.

Finding patient names with spelling or typographical errors is challenging. The only patient last name that NLM-NS missed was a complicated typographical error. Expanding all of the closely linked patient names into all of their one letter deletion and substitution patterns would not have found the one NLM-NS missed. Neamatullah and colleagues[14] also found that expanding name searches to simple spell checker errors was not helpful. Further, multiple typos may obscure the patient name sufficiently to prevent identification. The real name behind the

two typos on which NLM-NS failed could not be guessed by our project's name taggers, and based on Social Security name data, 150 000 people in the US had a name token with the same spell-check distance from the mangled name as did the real name. Finding the names of relatives—though they are rare—can also be a challenge. The HL7 messages we used as our targets for scrubbing did not happen to include HL7's next of kin segments (NK1), which carry the name(s) of closest relative(s). Those who plan to scrub reports carried by HL7 messages should take advantage of the NK1 segments to help identify relatives' names.

One way to assure the absence of all varieties of patient names in narrative reports would be for report authors to stop mentioning names of patients, and their relatives, in the body of clinical reports altogether. Radiology reports were exemplary in the almost total exclusion of patient names in their reports. Medical specialty groups should encourage their members to do the same; for example, say 'the patient,' rather than referring to the patient by name, to facilitate the use of the test report body for research and/or other review purposes with less privacy risk.

Acknowledgements We are grateful to Bert M Kestenbaum of the Office of the Chief Actuary at the Social Security Administration, who provided us frequencies of names of applicants for a social security number, and Dr Jon McKeeby, CIO, Clinical Center at NIH and his staff for their help in obtaining and interpreting the clinical data. We acknowledge contributions of Dr Chris Lu to the design and coding of VTT, which has been central to our annotation efforts. We also thank Drs. Lynette Hirschman and Samuel Bayer as well as John Aberdeen of MITRE, for their generous help and offers to test MIST on our study data.

Contributors MK and CJMD are guarantors of this study, responsible for the overall content, study design, and development, conducted the study, had access to the data, and controlled the decision to publish. ACB, ZAD, GD, and SO helped design and develop the de-identification tools described in the study, helped conduct the study, and generated the data. FMC contributed biostatistical expertise to the study design, and performed the data analysis.

Funding This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Disclaimer The findings and conclusions in this report are those of the authors and do not necessarily represent the official positions of the National Library of Medicine, the National Institutes of Health, or the Department of Health and Human Services.

Competing interests MK receives royalties from the University of Pittsburgh for his contributions to a de-identification project. The resulting product was acquired by a third party, which today is known as the De-ID Data Corp.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 U.S. Department of Health and Human Services. Public Welfare; Administrative Data Standards and Related Requirements; Security and Privacy; Privacy of Individually Identifiable Health Information; Other Requirements Relating to Uses and Disclosures of Protected Health Information. 45 C.F.R. Sect. 164.514. 2002. http://edocket.access.gpo.gov/cfr_2002/octqtr/pdf/45cfr164.514.pdf (accessed 18 Apr 2013).
2 Grishman R, Sundheim B. Message understanding conference-6: a brief history. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING'96). 5–9 Aug 1996, Copenhagen, Denmark, 1996:466–71.
3 Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 2011;12:417–28.
4 Deleger L, Molnar K, Savova G, *et al*. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J Am Med Inform Assoc* 2013;20:80–4.
5 Meystre S, Friedlin F, South B, *et al*. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol* 2010;10:70.
6 Mehnert RB. A world of knowledge for the nation's health: the U.S. National Library of Medicine. *Am J Hosp Pharm* 1986;43:2991–7.
7 McCray AT, Sponsler JL, Brylawski B, *et al*. The role of lexical knowledge in biomedical text understanding. In: Stead W. ed *Proceedings of the Eleventh Annual SCAMC; 1987*. IEEE Computer Society Press, 1987:103–7.
8 Schoolman HM, Lindberg DA. The information age in concept and practice at the National Library of Medicine. *Ann Am Acad Pol Soc Sci* 1988:117–26.
9 U.S. Department of Health and Human Services, Office of Civil Rights. Guidance on De-identification of Protected Health Information, 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html (accessed 18 Apr 2013).
10 Health Level Seven International. V2 Messages. http://www.hl7.org/implement/standards/v2messages.cfm (accessed 18 Apr 2013).
11 Henderson M. *HL7 Messaging*. 2nd edn. Aubry, Texas: Otech, 2007.
12 Brady K, Sriram R, Lide B, *et al*. Testing the Nation's Healthcare Information Infrastructure: NIST perspective. IEEE NIST. 2012 Nov; 0018-9162/12:50-7. http:/ComputingNow.computing.org
13 The H.I.S. *Desk Reference: A CIO Survey*. Baltimore, MD: CHIME and HCIA, Inc. 1998:26–9. ISBN 1-57372-033-X
14 Neamatullah I, Douglass M, Lehman LH, *et al*. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008;8:32.
15 Wellner B, Huyck M, Mardis S, *et al*. Rapidly retargetable approaches to de-identification in medical records. *J Am Med Inform Assoc* 2007;14:564–73.
16 Carpenter B. LingPipe for 99.99% recall of gene mentions. In: Proceedings of the 2nd BioCreative workshop; 23–25 April 2007, Madrid, Spain.
17 Cunningham H, Maynard D, Bontcheva K, *et al*. GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). 7–12 July 2002, Stroudsburg, PA, 2002.
18 Hopcroft JE, Ullman JD. *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley, 1979.
19 National Technical Information Service, U.S. Department of Commerce. Social Security Administration's Death Master File. http://www.ntis.gov/products/ssa-dmf.aspx (accessed 18 Apr 2013).
20 HHS Office of Civil Rights. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html
21 Minkov E, Wang RC, Tomasic A, *et al*. NER Systems that suit user's preferences: adjusting the recall-precision trade-off for entity extraction. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (ACL); NY. Jun 2006. 93–6. http://acl.ldc.upenn.edu/N/N06/N06-2024.pdf (accessed 6 Aug 2013).
22 Ye N, Chai KMA, Lee WS, *et al*. Optimizing F-measures: a tale of two approaches. Proceedings of the 29th Internat Conf on Machine Learning (ICML), Edinburgh, UK, 2012. http://icml.cc/2012/papers/175.pdf
23 Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 6* 1900;50:157–75.
24 The R Project for Statistical Computing. http://www.r-project.org/ (accessed 18 Apr 2013).
25 Beckwith B, Mahaadevan R, Balis U, *et al*. Development and evaluation of an open source software tool for de-identification of pathology reports. *BMC Med Inform Decis Mak* 2006;6:12.
26 Friedlin J, McDonald C. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc* 2008;15:601–10.
27 Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;14:550–63. Paper and data supplement available at: http://jamia.bmj.com/content/14/5/550.long