Computer Algorithm for Automated Work Group Classification From Free Text: The DREAM Technique

Philip Harber, MD, MPH
Lori Crawford, BS
Amarpreet Cheema, BS
Levanto Schacter, MS

Objective: *This study developed and tested a computer method to automatically assign subjects to aggregate work groups based on their free text work descriptions.* ***Methods:*** *The Double Root Extended Automated Matcher (DREAM) algorithm classifies individuals based on pairs of subjects' free text word roots in common with those of standard classification systems and several explicitly defined linkages between term roots and aggregates.* ***Results:*** *DREAM effectively analyzed free text from 5887 participants in a multisite chronic obstructive pulmonary disease prevention study (Lung Health Study). For a test set of 533 cases, DREAMs classifications compared favorably with those of a four-human panel. The humans rated the accuracy of DREAM as good or better in 80% of the test cases.* ***Conclusions:*** *Automated text interpretation is a promising tool for analyzing large data sets for applications in data mining, research, and surveillance. Work descriptive information is most useful when it can link an individual to aggregate entities that have occupational health relevance. Determining the appropriate group requires considerable expertise. This article describes a new method for making such assignments using a computer algorithm to reduce dependence on the limited number of occupational health experts. In addition, computer algorithms foster consistency of assignments.* (J Occup Environ Med. *2007;49:41–49*)

Traditional approaches using narrowly focused questionnaires are difficult to apply in community-based studies because work is commonly very diverse and because of frequent job changes. Often, a three-stage process is used: 1) a database is created with information about exposures typical of broad aggregate groups; 2) an individual is assigned to a work group; and 3) exposures typical of the work group are linked to the individual. For example, personnel records may be used to determine that an individual is a welder; then, a job exposure matrix that contains exposures typical of welders assigns exposures to the individual. The alternative approach of using subjects' free text work descriptions allows diverse input but usually requires complex job exposure matrices and expensive classification by professional experts.[1]

This article describes the development and initial evaluation of a computer-based algorithm—Double Root Extended Automated Matcher (DREAM)—for automatically interpreting free text to classify individuals into work groups to accomplish the second step. This approach is fundamentally different from methods used in other computer systems.

Materials and Methods

Data were derived from participants in the Lung Health Study, a multisite smoking cessation and bronchodilator administration trial for persons with early chronic obstructive pulmonary disease (COPD).[2,3] The

From the Division of Occupational and Environmental Medicine, Department of Family Medicine, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, California.

Supported by the American Association of Medical Colleges (AAMC), National Institute for Occupational Safety and Health (NIOSH)/Centers for Disease Control (CDC) #MM0060 02/02; National Cancer Institute (NCI) RO1 0H03839-01; and Contract N01-HR-46002 from the Division of Lung Diseases, National Heart, Lung, and Blood Institute, National Institutes of Health. (Lung Health Study).

Address correspondence to: Philip Harber, MD, MPH, Division of Occupational and Environmental Medicine, David Geffen School of Medicine at UCLA, 10880 Wilshire Boulevard, Suite 1800, Los Angeles, CA 90024; E-mail: pharber@mednet.ucla.edu

DOI: 10.1097/01.jom.0000251826.37828.2e

subject group was comprised of 5887 subjects who participated in 10 clinical centers in North America (nine in the United States, one in Canada).

The participants answered several occupational questions at baseline; these included three questions that allowed free text responses: job, industry or field, and title. These were asked for both "current/most recent" and "usual" occupation. Data described in this article are from the "current/most recent occupation" responses.

The responses to the three specific questions were concatenated (joined) for analysis purposes. A detailed description of the lexical analysis methodology can be found in another paper.[4] The lexical analyzer reduced words to their roots (morphemes) (eg, "nursing" → "nurs") and then created a list of roots and pairs of roots for each subject. A method to assign individual subjects to meaningful work aggregate groups was developed as follows.

Terms and numeric codes used in the North American Industrial Classification System (NAICS)[5] and Standard Occupational Codes (SOC)[6] systems were examined to develop ad hoc groupings based on continuous numeric codes. Using an iterative manual process, groups were modified and combined to create 61 work groups (WGs). The selection process was based on the investigators' knowledge of exposures suspected of association with lung disease. In addition, group selection was facilitated by analysis of terms in common in NAICS and SOC descriptions. The selection of the 61 groups was affected by the investigators' intention to use the method for an epidemiologic study of COPD. In addition, they considered statistical issues to guide the degree of "granularity" (ie, number of classification groups). Different aggregate groups might be selected by investigators evaluating musculoskeletal injuries with a much larger population size. The number of groupings is comparable to group numbers used in a study conducted by others.[7] By intent, these included both job title and industry content.[4]

Once these WGs were defined, an algorithm using four techniques for linking individuals to these groups was implemented:

1. Single root matches (1R): when the DREAM program noted that an input phrase has a single root in common with the SOC or NAICS descriptive text for the range of codes included in the WG, a link was created.
2. Double root matches (2R): when an input phrase has a pair of roots in common with the aggregate WG text, a link to the WG was created. The root pair was restricted to a lexical distance of ≤3 words apart in the subject's responses (ie, the two terms must be separated by ≤2 words).
   Use of only these two techniques was inadequate. Using the single root criterion, virtually all subjects (97%) linked to at least one WG, but many were nonspecific or inaccurate (eg, "treat" may be related to either Health Care or Metal Production). With double root matching, a large proportion of subjects (61%) did not match to any aggregates. Because of the nonspecificity of single root matches, these were not subsequently used. Therefore, these two techniques were extended:
3. Forced single root matches (F1R): common single roots of individuals without any 2R matches were examined to determine those that could be matched unambiguously based only on a single root (eg, an individual who used the term "baker" could be reasonably matched to the "food processing" WG on the basis of only a single root).
4. Forced double root matches (F2R): a similar approach to that of forced single root matches was used to create the forced double root matches based on the presence of a pair of roots in the subjects' text string. For example, the concomitant use of "president" + "vice" or "officer" + "correctional" allows mapping to a specific WG (management and law enforcement, respectively).

The overall method is termed DREAM; it incorporates 2R, F1R, and F2R-based linkages

By the DREAM technique, many subjects had links to several WGs. Therefore, assignment to WGs is expressed as two sets: "all links" (all possible WGs for which the subject matches) and "best link" (the single WG for the subject that has the greatest strength based on numbers of mappings or hierarchic value of match type: forced double root > forced single root > double root empiric). (In the event of a tie, selection of best is random). This approach is consistent with "fuzzy set theory"[8] because it assigns membership strength and supports multiple memberships.

Evaluation

The primary goal of this work was to develop a new method (ie, "proof of concept"). Thorough evaluation of the concept should be done in the future after optimization of the computer algorithms and using an independent data set. Nevertheless, we did perform limited assessments. The primary evaluation was assessment of ability to assign participants to WGs.

Two secondary evaluation methods were used: face validity and human panel. Face validity was assessed by reviewing the proportion of subjects in each WG who self-reported dust exposure, fume exposure, or use of a mask at work (using the best group). The degree to which this appeared to be generally known work characteristics provided a qualitative assessment (eg, dust exposure is a priori expected to be more common in construction than in education WGs).

The panel methods were based on the opinions of four humans. Two of

TABLE 1A
Human Validation of Computer-Assigned Work Groups

		Best		All (not Best)	
Range	Score	n	Percent	n	Percent
Very good	4.5–5.0	265	49.7	126	24.8
Good	3.5–4.49	154	28.9	120	23.6
Close	2.5–3.49	73	13.7	139	27.3
Possibly wrong	1.5–2.49	33	6.2	97	19.1
Bad	<1.5	8	1.5	27	5.3
	Mean	4.06		3.19	

The table shows the distribution of ratings of the accuracy of work group assignment by DREAM. Ratings are based on the average of the four raters for each case. χ^2 with test for trend showed that best and all differed with $P < 0.0001$.

TABLE 1B
Evaluation: Dream's Selection of Humans' Choice

	N	A: Human Best = DREAM Best	B: Human Best Included in DREAM's All	C: Not Linked by DREAM
Chosen as BEST by 4 humans	233	187	11	35
Chosen as BEST by 3 humans	162	108	8	46
Chosen as BEST by 2 humans	156	88	12	56
Chosen as BEST by 1 human	402	180	20	202

Four humans independently selected the optimal work group (WG) for the test cases. The table summarizes how frequently DREAM included the humans' choices: A, human and DREAM selected same WG; B, DREAM assigned the humans' choice, but not as DREAM's "best"; C, DREAM did not select the humans' choice.

TABLE 1C
Evaluation: Concordances of Assignments

Rater	2	3	4	DREAM
1	66.2%	60.2%	62.7%	58.3%
2		64.2%	68.3%	55.7%
3			61.2%	53.1%
4				53.5%

Percentage agreements of best work group assignments between pairs of human raters and DREAM and each rater are shown.

the humans were investigators (P.H., L.C.) who had familiarity with the computer principles used in the study; the other two were two general occupational health staff members without any prior knowledge of the DREAM system. Two subsets of data were randomly selected for evaluation. The first included only the best WGs selected by the DREAM system (533 randomly selected participants). The second was based on the all links selected by DREAM for 509 randomly selected participants; only one of the "all links," but not the "best," was randomly selected for each of the 509 evaluation cases. First, the humans rated the accuracy: Each was shown the WG assigned by DREAM and the subject's textual information. The humans rated the appropriateness of the WG assignment using a 5-point scale (see Table 1) The results from the four raters were averaged. This was done separately for "all" links and the single "best" group. Second, the humans were provided the participants' textual information and asked to select the most appropriate WGs. The WGs selected by the humans were then compared with those selected by the computer system. Concordance of humans and computer was determined in a pairwise fashion (ie, for four humans and one computer, there were 10 comparisons). In addition, human–DREAM agreement was compared with agreement of a 25% random sample of humans' assignments with the other 75% of human assignments.

Results

Table 2 summarizes the efficacy of linking subjects to WGs. Subjects' descriptions commonly (97%) had at least one root in common with WG terms derived from the standard classification schemes (NAICS and SOC). However, the word in common was frequently nonspecific and led to many obviously spurious matches. Only 40% of subjects had at least two word roots in common with the standard classifiers (2R), but the linkages appeared to be valid. The "forced root mappings," which depend on a limited number of des-

TABLE 2
Efficacy of Matching Algorithms: Number of Links per Subject

	Subjects With ≥1 Match		Number of Links/Case (for those with ≥1)					
Method	n	Percent	Average	Standard Deviation	Maximum	Q1	Median	Q3
Single root (1R)	5691	97%	12.3	7.24	42	7	11	17
Double root (2R)	2293	39%	1.8	1.2	10	1	1	2
F1R	5417	92%	2.4	1.3	9	1	2	3
F2R	706	12%	1.1	0.2	3	1	1	1
DREAM	5537	94%	3.2	2	15	2	3	4

Method refers to algorithm for creating links: F1R, forced 1 root; F2R, forced double root (see “Methods” section); DREAM includes 2R, F1R, F2R; Q1, Q3, first and third quartiles.

TABLE 3
Frequency of Assignment to Work Groups

Best Group						All Groups	
Men			Women			Men and Women Combined	
Group	n	Rank	Group	n	Rank	Group	n
Sales	293	1	Office	310	1	Office	1691
Construction	263	2	Finance	254	2	Professional	1620
Finance	240	3	Health care	220	3	Sales	1262
Professional	220	4	Home work	203	4	Finance	801
Repair and maintenance	199	5	Education	167	5	Management	749
Auto mechanic	175	6	Sales	158	6	Maintenance and cleaning	711
Education	134	7	Food service	94	7	Repair and maintenance	649
Transportation—truck	117	8	Professional	94	7	Construction	537
Office	108	9	Community services	44	9	Health care	436
Computer	104	10	Computer	40	10	Auto mechanic	429
Metal production	101	11	Legal	39	11	Education	417
Law enforcement	90	12	Maintenance and cleaning	36	12	Motor vehicle driver	367
Engineer architect	86	13	Management	31	13	Home work	332
Maintenance and cleaning	83	14	Personal appearance workers	27	14	Engineer architect	319
Postal services	77	15	Postal services	26	15	Metal products	297
Health care	74	16	Nursing	23	16	Food service	287
Management	71	17	Printing	23	16	Manufacturing	275
Food service	66	18	Construction	19	18	Computer	239
Metal products	64	19	Auto mechanic	18	19	Law enforcement	236
Agriculture	62	20	Entertainment and sports	18	19	Machinery	212
Legal	62	20	Food processing	17	21	Metal production	206
Transportation–not otherwise specified	61	22	Law enforcement	17	21	Transportation–not otherwise specified	203
Machinery	57	23	Repair and maintenance	17	23	Food processing	196
Firefighter	50	24	Machinery	15	24	Transportation—truck	186
Printing	46	25	Manufacturing	15	24	Community services	172

ignated mappings, matched 92% of subjects by F1R and 12% by the presence of two common roots (F2R). Overall, the DREAM system found at least one match for >94% of subjects. Most subjects were matched to a small number of WGs; the median number of assignments per subject was 3.

Table 3 shows the frequency of assignments to the most common WGs. Results are shown for the single best group separately by gender and for all groups combining men and women. As shown by the best group data in Table 3, “white collar” work was most common, ie, office, finance, and sales were most common. The pattern of WG ranks by gender suggests that assignments were accurate. Construction (rank 2 in men vs 18 in women), auto mechanic (6 vs 19), and truck transportation (8 vs 26) were much more common in men than women, whereas education, health care, and nursing were much more common in women. Similar patterns were seen when considering “all links”; for example, construction (7 vs 20) and auto mechanic (8 vs 23) were common in men, whereas work in health care (22 vs 5) and nursing (55 vs 11) were

TABLE 4
Frequency of Self-Stated Exposures by Best Work Group

Work Group	N	Dust (%)	Fume (%)	Mask (%)
Agriculture	74	46	38	12
Auto mechanic	193	40	43	11
Community services	71	11	6	0
Computer	144	2	2	1
Construction	282	47	26	12
Education	301	12	5	2
Entertainment and sports	63	8	6	0
Finance	494	5	3	1
Firefighter	54	44	44	39
Food service	160	6	13	1
Health care	294	7	9	4
Home work	225	8	6	1
Legal	101	4	2	2
Management	102	7	6	0
Metal production	111	50	44	13
Metal products	68	59	60	21
Office	418	8	4	0
Postal services	103	36	5	0
Printing	69	28	41	0
Professional	314	7	7	2
Sales	451	8	6	1
Personal appearance workers	38	11	61	3

Numbers in dust, fume, and mask columns refer to the percentage of persons in each work group who reported dust exposure, fume exposure, or mask use, respectively. Work groups shown are the most common.

much more common in women. Less common WGs not shown in the table are appliance manufacturing, artist, chemical production, computer and electronics manufacturing, furniture, hotel, laundry, military, mining, motor vehicle driver, motor vehicle sales, paper, personal service, petroleum, plastics, scientist, textile, tourism, transportation–air, transportation–rail, warehouse, wholesale trade, wood, and not working.

Evaluation

Face validity was assessed by examining the relationship of WG assignment to self-reported exposures for groups commonly considered to have high or low exposures. The proportion of members of the common WGs who reported dust exposure, fume exposure, or mask use is shown in Table 4. Groups with particularly frequent self-reported dust exposure include: agriculture (46%), construction (47%), metal products (59%), and metal production (50%). Self-reported dust exposure was infrequent in professionals (7%), sales (8%), and computer workers (2%). Mask use was particularly common in firefighters, 39% of whom reported using masks.

Table 1 summarizes the evaluation of the computer-based matches using a random validation subsample. Rating of the accuracy of matches is shown in Table 1A. Overall, 95% of the best and 76% of the all matches were in an acceptable range as determined by the human raters, and only 1.5% of the best matches were considered definitely wrong.

Humans identified their choices for the best WG based on review of the free text responses of the test set cases. Table 1B describes how frequently the humans' best WG assignments were included among those selected by the computer. For 233 of the 533 test cases, all four humans selected the same WG for the "Best" assignment. The computer agreed about the best assignment in 187 of these 233 instances. For 11 of the remaining 46, the computer included the humans' selection in its "all" matches if not selected as the single best. Thus, the computer failed to include an assignment unanimously selected by humans in only 35 of 233 instances.

Figure 1 illustrates the types of DREAM links (ie, DREAM best, DREAM all, not linked) according to the number of humans selecting a WG assignment.

If DREAM performed comparably to humans, one would expect that DREAM would be more likely to select a WG if all humans selected it than if only one human chose the WG. The data confirm this gradient, noting the degree of human–computer concordance (% DREAM Best) is related to the degree of human–human concordance. (For example, DREAM agreed with the humans in 80% of instances when all 4 humans agreed about the best choice, but when only two humans agreed, DREAM made the same choice in only 56% of instances.)

The degree of agreement between humans and DREAM was comparable to the agreement among humans (Table 1). The four humans agreed on the best assignment in only 44% of the test cases (233 of 533). Table 1C shows the concordance between pairs of humans and human–computer pairs. Pairs of humans agreed between 60% and 68% of cases, and agreement between computer and human pairs was from 53% to 58%. In general, the humans used a more narrow range of WGs than did the computer.

An additional evaluation was conducted limited to those study participants for whom the humans raters were reasonably consistent (ie, either three or four of the humans concurred about the best choice of WG). For this group, an evaluation set of raters' assignments was created by randomly selecting 25% of the humans' assignments (random rater evaluation set). Then, the agreement of DREAM with this random set of humans' assignments was determined; similarly, the agreement of the random evaluation set with the

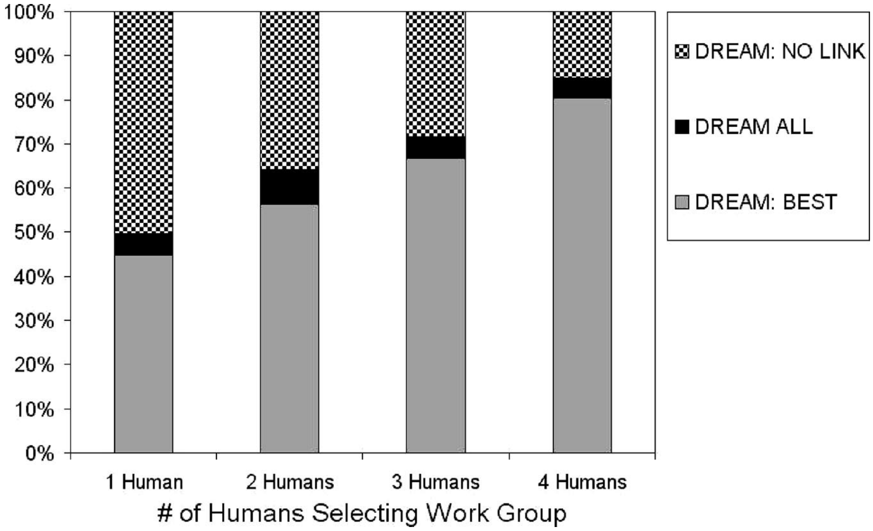

Fig. 1. DREAM linkages according to consistency of humans' work group assignments. Work group assignments are classed by consistency of humans (number of human raters selecting the work group). The figure shows the consistency of DREAM links within each class. DREAM and humans showed most consistency when the humans themselves were consistent.

other humans was determined. Where DREAM was compared with the random evaluation set, there was 66% concurrence (based on a sample of 402). There was 79% agreement between the random evaluation set and the other humans.

Then, a set of "outlier human assignments" was created; for those study participants for whom three of the raters concurred, the outlier human assignments were those selected by only one rater. DREAM agreed with the outlier human rater in only 17% of the 174 instances.

Discussion

Occupational health practice and research often require classifying workers into broad aggregate WGs, generally reflecting similarities of exposure. This article describes a new approach based on computer analysis of free text work descriptions provided by subjects.

Several methods have been used for assigning individuals to aggregate occupational groups. In some, questionnaires have explicit links to exposure relevant aggregate groupings (eg, "Have you been exposed to asbestos? Are you a welder?"). Providing lists from which to choose is less feasible in community-based studies because of the wide range of agents, occupations, and industries. Increasingly, subjects provide free text descriptions of their work that are used to classify them into broader aggregate WGs. Grouping is often based on standard numeric codes (eg, SOC)[6] or based on sets of job/industry terms. Then, a mapping aggregation scheme is superimposed on the numeric codes. In some studies, subjects' text descriptions are linked to groups by "expert opinion" by one or more "domain experts" (such as an industrial hygienist or an occupational physician) who also express a semiquantitative opinion about the extent of exposure to a particular agent or to classify the subject into one of several large aggregate groups. These techniques depend on the availability of well training occupational health professionals. In the future, computer methods may overcome this limitation.

Siemiatycki et al assessed methods of classifying subjects by human experts according to exposures.[1] Although effective, this approach is expensive and may be difficult to generalize. They included only 160 discrete exposures, but general population-based studies are likely to require many more. Stewart reviewed several additional effective expert based approaches.[9] Mannetje's review explicitly considers utilization of computer-assisted classification methods.[10]

The DREAM method has several differences from these approaches. Once implemented, the DREAM technique does not require high-level human expertise. DREAM is therefore more economic and may be scaled up more readily than traditional classifications by human experts.[1,9] Furthermore, the method may be easily applied to develop new population specific job-exposure matrices, which generally perform better than general purpose classification schemes.[10–12] A computer-based system consistently applies the programmed algorithm, whereas individual human raters have inherent variability. For example, all four humans concurred about the "best" WG in less than half of the evaluation test cases. The computer algorithm is fully explicit, whereas human experts make assignments by an implicit subjective process ("expert judgment").

Computer classification based on free text permits flexibility of use of data. Data collected for one purpose can be used for other purposes in the future. For example, in the study illustrated here, data were collected for a smoking cessation trial but used for occupational health studies.[13,14]

The DREAM system integrates classification by job, industry, agent, and task. In the traditional approach, these are considered to be largely independent domains.[15] In a study assessing a worker lexicon, we found considerable overlap in the actual information domain content of terms used by workers. The parallel coding schemes were not developed for health purposes but rather for economic classification. Several investigators have also noted that simultaneously using information from several domains could improve precision[9,16]

Natural Language Processing

Natural language processing (NLP) techniques are computer systems for interpreting language. Natural language is extremely complex, including hundreds of thousands of words and highly variable constructions. Systems focused within a specific domain (eg, pathology classification[17,18]) are typically more successful than those attempting to handle all language.[19–23] An extremely large vocabulary is necessary to encompass all medicine,[24] but occupational health is a more restricted domain and fewer terms are necessary. (For example, 589 single roots accounted for 80% of actual terms used by the subjects.) The occupational health context helps reduce ambiguity, because the meaning of a particular term often depends on the context.

Evaluation

The effectiveness of the automated system (DREAM) was measured in several ways: Efficacy: 94% of subjects could be matched to one or more categories. This contrasts to the more limited success in finding a link with systems that use questionnaire modules for occupational exposure classification. For example, Stewart et al used approximately 38 modules to obtain approximately 40% classification coverage.[25,26]

A sample of participants' data was used for a secondary evaluation; results are summarized in Table 1. Accuracy of DREAM's assignments was by considered acceptable by human evaluators in most instances (Table 1A). WG assignments made by humans were compared with those made by the computer in two ways. Overall, DREAM selected the same best WG in 80% of the cases in which all four humans concurred. This compares favorably to human performance; all humans concurred in only 44% of the cases. In the random rater evaluation subset, DREAM human agreement was then compared with agreement between the randomly selected subset of humans and others (limited to study participants for whom at least three humans concurred). Although humans were generally more consistent with other humans (79% agreement) than DREAM with humans (66% agreement), the magnitude of difference was relatively small. Where humans themselves disagreed, DREAM generally agreed with the majority human opinion rather than with the outlier human; DREAM agreed with the humans' majority in 83% of the instances in which one human was an outlier. Face validity was suggested by the frequency with which subjects in each WG reported dust exposure, fume exposure, or mask use. Groups reporting frequent dust exposure (eg, construction, agriculture, metal products) and those reporting infrequent dust exposures (eg, computer group, professional, sales) are consistent with a priori expectations.

Limitations

The computer algorithm used here is relatively simple in comparison to methods used in other settings. Because the DREAM algorithms are less sophisticated than techniques used in many medical NLP applications, there is considerable opportunity for improvement in the future. DREAM is, however, an advance over simple lists of synonyms used in the past.[25]

WG assignment by computerized free text analysis will never be perfect. Occasional misunderstandings are due to inherent ambiguities in language. For example, "nursery" is inherently nonspecific; it might represent where plants are cultivated or part of the pediatric unit of a hospital. Indeed, such "double meanings" form the basis of much humor. In the future, context-dependent interpretation would improve performance. Furthermore, although DREAM was applied here in an "offline" (batch processing) mode, similar methods may be used iteratively in which subjects are asked to clarify any ambiguities.

Although the DREAM method is effective for assigning individuals to broad WGs, it does not in itself link exposures to individuals. Existing methods such as job exposure matrices must be used to summarize exposures typical of WGs. Thus, DREAM cannot completely replace existing methods, but rather, computer text analysis can improve one of the time-demanding steps in the overall process.

The evaluation sample was randomly chosen to avoid biases. However, the sample cases were selected from the same population used to optimize the coding and assignment methods. This might potentially favorably bias the evaluation. For example, if another population frequently used a term that was rarely mentioned by our study population, it would not have been selected for a forced match (F1R, F2R). Nevertheless, our study population was large and obtained from 10 geographically distinct regions. Furthermore, the methodology for implementing DREAM would also be applicable to the other population (ie, the same developmental technique could be applied to an entirely new population).

Future

The study shows that automated text coding for assigning to aggregate WGs may be a promising technology. More complex systems and further refinement will improve the performance in the future. In the future, automated coding methods may allow access to very large existing databases (eg, electronic medical records) and facilitate crossnational studies.[27,28] Currently, discovery of new exposure–health outcome relationships is often ad hoc and serendipitous; computerized methods may facilitate systematic "data mining."[29,30]

In addition to applications for research purposes, an automated classifier has direct applications in both public health and clinical practice settings. Identifying the broad WG is the first step in providing advice about medical assessment procedures, specific preventive interventions, and so on. There is a dearth of available public health expertise in most primary care and specialty settings, although a large number of the at-risk individuals are seen in these locations.[31,32] Computer-assisted knowledge delivery may help fill this gap.

Acknowledgments

The authors acknowledge the assistance of Drs. Donald Tashkin, Eva Hnizdo, John Connett, Michael Simmons, Katie Liu, and Jennifer Okada.

The principal investigators and senior staff of the clinical and coordinating centers, the NHLBI, and members of the Safety and Data Monitoring Board of the Lung Health Study are as follows: Case Western Reserve University, Cleveland, Ohio: M. D. Altose, MD (Principal Investigator), C. D. Deitz, PhD (Project Coordinator); Henry Ford Hospital, Detroit, Michigan: M. S. Eichenhorn, MD (Principal Investigator), K. J. Braden, AAS (Project Coordinator), P. A. Fantuz, RN, BSN, R. L. Jentons, MALLP (Project Coordinator); Johns Hopkins University School of Medicine, Baltimore, Maryland: R. A. Wise, MD (Principal Investigator), C. S. Rand, PhD (Coprincipal Investigator), K. A. Schiller (Project Coordinator); Mayo Clinic, Rochester, Minnesota: P. D. Scanlon, MD (Principal Investigator), G. M. Caron (Project Coordinator), K. S. Mieras, L. C. Walters; Oregon Health Sciences University, Portland, Oregon: A. S. Buist, MD (Principal Investigator), L. R. Johnson, PhD (LHS Pulmonary Function Coordinator), V. J. Bortz (Project Coordinator); University of Alabama at Birmingham, Alabama: W. C. Bailey, MD (Principal Investigator), L. B. Gerald, PhD, MSPH (Project Coordinator); University of California, Los Angeles, California: D. P. Tashkin, MD (Principal Investigator), I. P. Zuniga (Project Coordinator); University of Manitoba, Winnipeg, Canada: N. R. Anthonisen, MD (Principal Investigator, Steering Committee Chair), J. Manfreda, MD (Coprincipal Investigator), R. P. Murray, PhD (Coprincipal Investigator), S. C. Rempel-Rossum (Project Coordinator); University of Minnesota Coordinating Center, Minneapolis, Minnesota: J. E. Connett, PhD (Principal Investigator), P. G. Lindgren, MS, M. A. Skeans, MS, H. T. Voelker; University of Pittsburgh, Pittsburgh, Pennsylvania: R. M. Rogers, MD (Principal Investigator), M. E. Pusateri (Project Coordinator); University of Utah, Salt Lake City, Utah: R. E. Kanner, MD (Principal Investigator), G. M. Villegas (Project Coordinator); Data Monitoring Board: C. Furberg, MD, PhD, J. R. Landis, PhD, E. Mauger, PhD, J. R. Maurer, MD, Y. Phillips, MD, J. K. Stoller, MD, I. Tager, MD, A. Thomas, Jr., MD; Morbidity and Mortality Review Board: T. E. Cuddy, MD, R. S. Fontana, MD, R. E. Hyatt, MD, C. T. Lambrew, MD, B. A. Mason, MD, D. M. Mintzer, MD, R. B. Wray, MD; National Heart, Lung and Blood Institute Staff, Bethesda, Maryland: T. Croxton, MD, PhD (Project Officer), J. P. Kiley, PhD (Director, Division of Lung Diseases), G. Weinmann, MD (Director, Airway Biology and Disease Program, DLD), M. C. Wu, PhD (Division of Epidemiology & Clinical Applications).

References

1. Siemiatycki J, Dewar R, Richardson L. Costs and statistical power associated with five methods of collecting occupation exposure information for population-based case-control studies. *Am J Epidemiol*. 1989;130:1236–1246.
2. Buist AS, Connett JE, Miller RD, Kanner RE, Owens GR, Voelker HT. Chronic Obstructive Pulmonary Disease Early Intervention Trial (Lung Health Study). Baseline characteristics of randomized participants. *Chest*. 1993;103:1863–1872.
3. Connett JE, Kusek JW, Bailey WC, O'Hara P, Wu M. Design of the Lung Health Study: a randomized clinical trial of early intervention for chronic obstructive pulmonary disease. *Control Clin Trials*. 1993;14:3S–19S.
4. Harber P, Crawford L, Liu K, Schacter L. Working words: real-life lexicon of North American workers. *J Occup Environ Med*. 2005;47:859–864.
5. Office of Management and Budget. *North American Industry Classification System: United States, 1997.* Washington, DC: Executive Office of the President: Office of Management and Budget; US Government Printing Office, Supt. of Documents; 1998.
6. Office of Federal Statistical Policy and Standards. *Standard Occupational Classification Manual.* Springfield, VA: US Department of Commerce Technology Administration National Technical Information Service; 2000.
7. Hnizdo E, Sullivan PA, Bang KM, Wagner G. Airflow obstruction attributable to work in industry and occupation among US race/ethnic groups: a study of NHANES III data. *Am J Ind Med*. 2004; 46:126–135.
8. Steimann F. On the use and usefulness of fuzzy sets in medical AI. *Artif Intell Med*. 2001;21:131–137.
9. Stewart PA, Stewart WF. Occupational case–control studies: II. Recommendations for exposure assessment. *Am J Ind Med*. 1994;26:313–326.
10. Mannetje A, Kromhout H. The use of occupation and industry classifications in general population studies. *Int J Epidemiol*. 2003;32:419–428.
11. Kromhout H, Heederik D, Dalderup LM, Kromhout D. Performance of two general job-exposure matrices in a study of lung cancer morbidity in the Zutphen cohort. *Am J Epidemiol*. 1992;136:698–711.
12. Le Moual N, Bakke P, Orlowski E, et al. Performance of population specific job exposure matrices (JEMs): European collaborative analyses on occupational risk factors for chronic obstructive pulmonary disease with job exposure matrices (ECOJEM). *Occup Environ Med*. 2000; 57:126–132.
13. Harber P, Simmons M, Tashkin D, Hnizdo E, Crawford L, Connett J Comparison of 3 methods for assigning exposure metrics for evaluating occupation effect upon COPD. *Am J Repir Crit Care Med*. 2005;171:A818.
14. Harber P, Simmons M, Tashkin DP, Hnizdo E, Schachter L. What do 'dust,' fume,' 'mask use' really mean? *Am J Respir Crit Care Med*. 2003;167:506.
15. Stewart WF, Stewart PA. Occupational case–control studies: I. Collecting information on work histories and work-related exposures. *Am J Ind Med*. 1994; 26:297–312.
16. Burkhart G, Schulte PA, Robinson C, Sieber WK, Vossenas P, Ringen K. Job tasks, potential exposures, and health risks of laborers employed in the construction industry. *Am J Ind Med*. 1993; 24:413–425.
17. Sinha U, Yaghmai A, Thompson L, et al. Evaluation of SNOMED3.5 in representing concepts in chest radiology reports: integration of a SNOMED mapper with a radiology reporting workstation. *Proc AMIA Symp*. 2000;799–803.
18. Berman JJ, Moore GW. SNOMED-encoded surgical pathology databases: a tool for epidemiologic investigation. *Mod Pathol*. 1996;9:944–950.
19. Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. *J Biomed Inform*. 2001;34:4–14.
20. Baud R, Lovis C, Rassinoux AM, Michel PA, Scherrer JR. Automatic extraction of linguistic knowledge from an international classification. *Medinfo*. 1998;9: 581–585.
21. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. *J Am Med Inform Assoc*. 1999;6:143–150.
22. Huske-Kraus D. Text generation in clinical medicine—a review. *Methods Inf Med*. 2003;42:51–60.
23. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and

the representation of clinical data. *J Am Med Inform Assoc*. 1994;1:142–160.

24. Gu H, Perl Y, Geller J, Halper M, Singh M. A methodology for partitioning a vocabulary hierarchy into trees. *Artif Intell Med*. 1999;15:77–98.
25. Stewart PA, Stewart WF, Heineman EF, Dosemeci M, Linet M, Inskip PD. A novel approach to data collection in a case–control study of cancer and occupational exposures. *Int J Epidemiol*. 1996; 25:744–752.
26. Stewart PA, Stewart WF, Siemiatycki J, Heineman EF, Dosemeci M. Questionnaires for collecting detailed occupational information for community-based case control studies. *Am Ind Hyg Assoc J*. 1998;59:39–44.
27. Reynoso GA, March AD, Berra CM, et al. Development of the Spanish version of the Systematized Nomenclature of Medicine: methodology and main issues. *Proc AMIA Symp*. 2000;694–698.
28. Trombert-Paviot B, Rodrigues JM, Rogers JE, et al. GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. *Int J Med Inf*. 2000;58–59: 71–85.
29. Krauthammer M, Hripcsak G. A knowledge model for the interpretation and visualization of NLP-parsed discharged summaries. *Proc AMIA Symp*. 2001: 339–343.
30. de Bruijn B, Martin J. Getting to the (c)ore of knowledge: mining biomedical literature. *Int J Med Inf*. 2002;67:7–18.
31. Harber P. Primary care role in preventing occupational and environmental respiratory disease. *Prim Care*. 1994;21:291–311.
32. Harber P, Merz B. Time and knowledge barriers to recognizing occupational disease. *J Occup Environ Med*. 2001;43: 285–288.