WORD AND OBJECT IN DISEASE DESCRIPTIONS*

M.S. Blois, D.D. Sherertz, M.S. Tuttle
Section on Medical Information Science
University of California, San Francisco

Experiments were conducted on a book, Current Medical Information and Terminology, (AMA, Chicago, 1971, edited by Burgess Gordon, M.D.), which is a compendium of 3262 diseases, each of which is defined by a collection of attributes. The original purpose of the book was to introduce a standard nomenclature of disease names, and the attributes are organized in conventional medical form: a definition consists of a brief description of the relevant symptoms, signs, laboratory findings, and the like. Each disease is, in addition, assigned to one (or at most two) of eleven disease categories which enumerate physiological systems (skin, respiratory, cardiovascular, etc.). While the editorial style of the book is highly telegraphic, with many attributes being expressed as single words, it is nevertheless easily readable (see Figure 1).

The vocabulary employed consists of about 19,000 distinct "words" (determined by a lexical definition), roughly divided equally between common English words and medical terms. We measured word frequency by "disease occurrence", (the number of disease definitions in which a given word occurs one or more times). By this measure, only seven words occurred in more than half the disease definitions, and about 40% of the vocabulary occurred in only a single disease definition. (Table 1 lists the words at the top of the frequency list together with the number of occurrences.)

Assisted by the facilities of the TM UNIX operating system, we created a series of inverted files (from a magnetic tape of the CMIT text), and developed a set of interactive programs to form a word-and-context query system. This system has enabled us to study the problem of inferring term reference in this large sample of text (some 333,000 word occurrences), within the context of diseases.

An interesting early result was the ease with which many medical terms could be algorithmically separated from common English words. After adjusting for the fact that some disease categories are larger than others, we defined an entropy-like measure of the distribution of word occurrences over the eleven physiological categories as a measure of category specificity. We reasoned that some medical terms such as 'murmur', while not specific to any particular heart disease, are specific to heart disease generally. This term would not, for example, be used in describing endocrine disorders. Such a word would be expected to occur in category 04 (cardiovascular disease) frequently, and not in the other categories. Such a term would, by our measure, have a low 'entropy'. A common English word like 'of', would be used in the descriptions of all kinds of disease, and would accordingly have a high 'entropy'. Tables 2 and 3 show the top and bottom of the list of all words occurring in two or more diseases sorted by this entropy measure. In these lists, as our hypothesis seems to imply, low 'entropy' corresponds to high 'specificity', and high 'entropy' to low 'specificity'. This separation of medical terms from common English words, by algorithmic means, is facilitated by the context supplied by the notion of 'disease category', and the fact that this was represented in the CMIT text.

---

* This work was supported in part by grants from The Commonwealth Fund, and from the National Library of Medicine (1 K10 LM00014).

TM UNIX is a trademark of Bell Laboratories.

Our second experiment investigated the co-occurrence properties of some medical terms. Aware that many medical diagnostic programs have assumed attribute independence, we sought to shed light on the appropriateness of the assumption by evaluating it in terms of word co-occurrence in disease definitions.

Since the previously described procedure had given us a means of selecting medical terms from common English words, it was possible to produce lists of 'pure' medical terms. We then wrote a program which formed all pairs of such terms (ignoring order). We defined an 'association measure' (A) which measured the difference between the observed co-occurrences of term-pairs (they could co-occur in any location in the definition and in either order), and the co-occurrences expected from chance alone. Tables 4 and 5 show the top and bottom of a list of all pairs formed from the low entropy terms in the previous experiment. The first 1120 terms were chosen, that is, those having an entropy of 2.0 napiers or less. The pair list was then sorted by this association measure, A.

Word pairs which are found to be highly associated, appear to do so for two reasons. The test, which is trivial, is that some word pairs are semantically one word despite their being lexically, two. Common examples would be 'White House' and 'Hong Kong'; medical examples are 'vital capacity', 'axis deviation', and 'slit lamp'. These could have been avoided algorithmically by not taking adjacent words in forming the term-pairs, without any significant overall effect. The second reasons for high frequency word co-occurrence is that both words are causally related through underlying physiological mechanisms. It is these which had the greatest interest for us, and the measure A, may be viewed as a measure of the non-independence of the symptoms or signs themselves.

The term pairs which are negatively associated, have this property for the same reason. If the two terms are used typically in the descriptions of different diseases, they are less likely to co-occur than by chance. (In a baseball story on the sports page, we would not find 'pass', 'punt', or 'tackle'). These negatively associated pairs may have value in diagnostic programs for the recognition of two or more diseases in a given patient, a problem not satisfactorily dealt with by even the most sophisticated of current programs.

Finally, an extension of the entropy concept permits one to generate (algorithmically) the vocabularies used by the medical specialties (which correspond to the disease categories represented in CMIT. This is done by assigning terms which occur predominantly in one category to a single vocabulary and then sorting by entropy. Tables 6 and 7 show the vocabularies used in dermatology and gastroenterology (as derived from CMIT). These vocabularies, it will be noted, can be used as 'hit lists' for the purpose of recognizing the content of medical texts.

In summary, we see the ability to differentiate medical terms from common words by context, and the ability to relate the medical words by meaning, as two of the first steps toward text processing algorithms that preserve and can manipulate the semantic content of words in medical texts.

COLORADO TICK FEVER 00 2217

AT FEVER, MOUNTAIN; FEVER, MOUNTAIN TICK.

ET VIRUS TRANSMITTED BY TICK DERMACENTOR ANDERSONI.

SM CHILLS; HEADACHE; PHOTOPHOBIA; BACKACHE; PAIN IN EYE; MYALGIA; ANOREXIA; NAUSEA; VOMITING; PROSTRATION.

SG SEASONAL, MARCH TO JULY, IN WESTERN UNITED STATES; INCUBATION PERIOD 4-6 DAYS; ONSET ABRUPT; POSSIBLY SLIGHT ERYTHEMA; SUSTAINED FEVER, 102-104 F OR HIGHER SIGNIFICANT; PULSE RATE INCREASED. COURSE: IN PREVENTION, REMOVAL OF TICK FROM SKIN; APPLICATIONS TO SKIN OF TURPENTINE, IODINE, ACETONE; REMOVAL OF TICK BY INSERTION OF NEEDLE BETWEEN MOUTH PARTS; ASPIRIN FOR PAIN; ANTIBIOTIC TREATMENT INEFFECTIVE.

CM ENCEPHALITIS, MENINGITIS ESPECIALLY IN CHILDREN.

LB WBC DECREASED: MONOCYTOSIS; COMPLEMENT-FIXATION TEST POSITIVE; INJECTION OF SERUM OR CSF KILLING SUCKLING MICE; NEUTRALIZATION OF VIRUS WITH IMMUNE SERUM RESULTING IN SURVIVAL.

Figure 1. Typical disease 'definition' taken from CMIT

3206 of	507 small	364 other
2865 in	492 possible	360 acute
2405 possibly	489 severe	360 years
2315 with	478 most	358 failure
2104 course	473 disease	349 between
2010 to	457 pressure	349 large
1953 or	447 absence	341 dyspnea
1408 by	446 trauma	341 early
1379 usually	443 chronic	340 weakness
1194 pain	442 edema	339 nausea
980 as	435 percent	338 tenderness
945 on	434 treatment	337 inflammation
889 from	432 vomiting	337 mass
812 infection	431 later	336 age
766 features	426 absent	335 within
749 unknown	422 common	332 if
738 at	421 asymptomatic	331 lower
716 cells	420 during	328 swelling
699 associated	415 rarely	327 necrosis
682 increased	414 hereditary	325 positive
674 onset	401 lesions	324 headache
666 tissue	396 than	318 frequent
650 blood	390 abdominal	316 wbc
627 normal	389 more	315 area
619 skin	389 often	313 hemorrhage
603 and	383 into	313 infiltration
596 for	382 type	309 obstruction
575 rare	381 bone	304 form
553 fever	375 involvement	301 congenital
541 loss	369 especially	301 enlargement
538 after	367 areas	301 progressive

Table 1. The highest frequency words used in CMIT, together with the number of disease definitions in which the word occurs at least once.

0.8762	.02	.02	.02	.02	.01	.02	.01	.01	.03	.02	.82	lens 50
1.0204	.05	.04	.01	.02	.02	.03	.02	.00	.02	.01	.77	cornea 79
1.0278	.02	.01	.02	.02	.04	.04	.02	.01	.78	.02	.03	pbi 18
1.0369	.04	.11	.01	.77	.04	.02	.01	.01	.03	.04	.02	expectoration 54
1.0377	.02	.01	.03	.01	.74	.11	.02	.01	.02	.01	.02	murmur 86
1.0411	.03	.01	.01	.02	.77	.03	.02	.01	.04	.02	.03	atrium 45
1.0422	.04	.01	.01	.02	.02	.03	.02	.01	.04	.03	.77	ciliary 33
1.0656	.03	.04	.02	.02	.02	.02	.01	.01	.03	.03	.76	iris 53
1.0667	.01	.01	.01	.02	.03	.03	.01	.76	.02	.05		uptake 20
1.0885	.05	.01	.01	.02	.03	.03	.01	.02	.76	.03	.02	iodine 27
1.1236	.08	.01	.01	.03	.02	.74	.03	.01	.03	.01	.03	hemoglobin 51
1.1298	.01	.01	.02	.05	.74	.04	.02	.02	.05	.02	.03	qrs 41
1.1361	.02	.01	.02	.01	.69	.18	.02	.02	.02	.01	.01	systolic 91
1.1465	.03	.03	.03	.02	.02	.02	.02	.01	.03	.04	.74	glaucoma 49
1.1504	.07	.01	.01	.72	.09	.02	.02	.02	.02	.01	.01	rales 96
1.1556	.05	.00	.02	.05	.70	.01	.01	.01	.12	.02	.01	ecg 150
1.1642	.01	.02	.02	.73	.04	.04	.02	.01	.05	.02	.03	bronchoscopy 26
1.1716	.07	.02	.06	.02	.01	.02	.01	.01	.03	.02	.72	cataract 53
1.1812	.02	.03	.01	.03	.04	.04	.01	.01	.03	.07	.72	retina 61
1.1959	.01	.03	.03	.03	.02	.04	.02	.73	.05	.02	.03	urethral 64
1.2069	.01	.03	.02	.03	.02	.04	.02	.72	.05	.02	.03	urethra 58
1.2106	.01	.02	.02	.03	.02	.04	.02	.72	.04	.02	.04	cystoscopy 53
1.2124	.03	.02	.02	.03	.02	.04	.02	.01	.05	.04	.72	vitreous 24
1.2262	.02	.70	.01	.02	.02	.02	.02	.01	.03	.01	.06	epidermis 93
1.2293	.03	.01	.02	.03	.04	.04	.02	.72	.05	.02	.03	cervix 64
1.2493	.04	.01	.03	.02	.70	.05	.01	.01	.04	.03	.02	atrial 65
1.2551	.05	.01	.02	.02	.04	.01	.01	.01	.06	.08	.69	vision 192
1.2560	.02	.03	.02	.03	.04	.04	.02	.01	.05	.02	.71	intraocular 23
1.2583	.04	.01	.02	.05	.02	.04	.02	.70	.05	.02	.03	pyuria 65
1.2689	.04	.01	.04	.04	.01	.05	.01	.69	.07	.01	.02	uterus 96
1.2758	.01	.02	.02	.04	.69	.05	.02	.01	.04	.02	.04	angiocardiography 30
1.2913	.01	.01	.02	.05	.64	.03	.01	.01	.07	.14	.01	ventricle 95
1.3008	.01	.04	.02	.05	.04	.04	.03	.03	.69	.02	.03	adenoma 21
1.3089	.01	.68	.02	.02	.01	.02	.04	.04	.06	.04	.04	dermis 35
1.3092	.01	.03	.04	.02	.02	.05	.03	.07	.69	.03	.02	hormone 44
1.3108	.09	.01	.01	.67	.03	.03	.05	.02	.04	.01	.02	sputum 57
1.3120	.02	.01	.02	.08	.67	.07	.02	.01	.05	.02	.03	mitral 38
1.3176	.05	.01	.03	.02	.04	.08	.02	.68	.03	.01	.02	uterine 98
1.3185	.04	.02	.02	.69	.02	.05	.02	.03	.06	.02	.04	alveoli 24
1.3190	.03	.06	.02	.01	.01	.02	.02	.07	.68	.03	.05	pituitary 52
1.3211	.03	.03	.01	.05	.67	.05	.01	.01	.07	.04	.02	aorta 60
1.3224	.05	.06	.03	.02	.04	.68	.02	.01	.05	.02	.03	splenectomy 27
1.3269	.02	.02	.02	.04	.03	.68	.03	.02	.07	.03	.05	target 11
1.3317	.02	.03	.04	.04	.04	.03	.04	.69	.04	.01	.02	vaginal 91
1.3344	.03	.02	.02	.03	.68	.05	.02	.03	.06	.02	.04	gallop 29
1.3369	.03	.02	.01	.06	.10	.03	.03	.01	.04	.01	.66	chamber 41
1.3375	.05	.02	.02	.03	.04	.04	.02	.02	.68	.02	.07	hyperglycemia 18
1.3378	.04	.03	.03	.04	.02	.02	.01	.00	.05	.08	.67	eye 113
1.3395	.04	.01	.01	.06	.65	.05	.01	.01	.08	.07	.01	ventricular 110
1.3439	.02	.03	.02	.02	.02	.04	.02	.01	.05	.13	.65	pupil 29
1.3560	.15	.03	.01	.01	.01	.02	.01	.01	.07	.08	.61	corneal 84
1.3585	.02	.02	.02	.02	.03	.15	.01	.04	.64	.02	.02	ber 28
1.3605	.02	.02	.02	.03	.65	.12	.04	.01	.05	.02	.03	valve 35
1.3690	.04	.01	.01	.02	.65	.03	.03	.01	.08	.08	.03	wave 47
1.3703	.01	.05	.02	.03	.02	.04	.02	.01	.11	.02	.65	lids 20
1.3785	.03	.02	.03	.65	.02	.04	.04	.01	.10	.02	.03	pneumothorax 28

Table 2. The lowest 'entropy' words in CMIT, in order of increasing 'entropy' The entropy is given in the first column; the entries in the next 11 columns are the percent of occurrences in the 11 disease categories (body as a whole, skin, musculo-skeletal, respiratory, cardiovascular, hemic and lymphatic, GI, GU, endocrine, nervous, organs of special sense).

2.3626	.08	.05	.08	.11	.13	.07	.06	.10	.11	.07	.13	degree	124
2.3629	.08	.06	.08	.09	.14	.13	.05	.07	.11	.09	.09	absent	126
2.3630	.09	.12	.09	.07	.11	.10	.05	.08	.13	.11	.08	bloody	8
2.3635	.05	.09	.11	.07	.10	.07	.10	.10	.13	.11	.07	common	422
2.3637	.12	.08	.09	.07	.05	.11	.10	.10	.14	.05	.09	china	4
2.3640	.10	.10	.10	.09	.10	.09	.09	.09	.05	.14	.06	within	335
2.3642	.08	.11	.12	.09	.08	.09	.13	.10	.06	.09	.05	marked	159
2.3647	.11	.08	.11	.06	.07	.13	.09	.05	.08	.10	.11	indicative	20
2.3653	.07	.09	.12	.05	.12	.09	.07	.10	.07	.11	.11	absence	447
2.3660	.09	.06	.08	.07	.09	.11	.09	.10	.09	.13	.12	milder	44
2.3667	.12	.06	.08	.13	.11	.07	.09	.10	.09	.06	.09	week	45
2.3668	.07	.09	.08	.06	.13	.10	.11	.10	.11	.09	.06	often	389
2.3678	.11	.05	.09	.10	.09	.07	.07	.11	.13	.07	.11	simple	46
2.3681	.12	.11	.06	.09	.08	.09	.07	.07	.14	.08	.09	2	130
2.3687	.09	.09	.09	.10	.08	.08	.09	.14	.12	.05	.07	large	349
2.3701	.08	.07	.13	.09	.13	.06	.09	.10	.07	.07	.11	causing	256
2.3708	.10	.06	.10	.10	.10	.14	.10	.06	.07	.09	.07	severe	489
2.3711	.06	.06	.12	.12	.10	.09	.11	.07	.11	.09	.09	late	125
2.3716	.09	.10	.11	.05	.13	.09	.08	.08	.06	.10	.11	without	246
2.3718	.09	.08	.08	.08	.11	.09	.13	.12	.09	.09	.05	if	332
2.3718	.10	.09	.09	.08	.13	.10	.05	.05	.11	.07	.09	increasing	123
2.3724	.13	.10	.09	.10	.05	.09	.07	.08	.11	.08	.09	for	596
2.3727	.07	.07	.13	.07	.11	.12	.08	.08	.11	.09	.08	than	396
2.3746	.08	.11	.10	.08	.08	.10	.11	.11	.08	.11	.06	most	478
2.3746	.07	.12	.13	.07	.10	.10	.10	.09	.06	.08	.08	each	30
2.3746	.09	.08	.09	.09	.07	.10	.08	.06	.11	.13	.09	onset	674
2.3748	.11	.09	.07	.08	.07	.11	.08	.08	.07	.10	.14	accumulation	61
2.3762	.09	.10	.07	.08	.11	.12	.07	.06	.11	.11	.07	poor	55
2.3776	.07	.11	.14	.08	.08	.09	.08	.08	.11	.10	.07	more	389
2.3780	.09	.09	.10	.12	.09	.08	.10	.11	.10	.06	.07	persistent	124
2.3783	.10	.09	.12	.05	.08	.10	.07	.09	.10	.11	.09	and	603
2.3792	.08	.09	.09	.10	.06	.11	.10	.09	.12	.09	.10	type	382
2.3793	.08	.09	.08	.07	.10	.10	.09	.10	.13	.09	.08	rarely	415
2.3794	.08	.08	.08	.08	.08	.12	.12	.10	.09	.11	.07	variable	205
2.3794	.09	.08	.08	.10	.12	.08	.09	.08	.08	.13	.08	cases	260
2.3801	.09	.09	.10	.07	.09	.12	.10	.12	.07	.08	.07	frequent	318
2.3815	.08	.10	.08	.11	.09	.08	.07	.08	.11	.11	.11	later	431
2.3819	.08	.08	.10	.11	.12	.09	.08	.10	.10	.09	.06	during	420
2.3821	.07	.10	.10	.11	.11	.08	.10	.06	.10	.09	.08	especially	369
2.3847	.06	.11	.11	.08	.09	.08	.09	.11	.08	.10	.10	usually	1379
2.3847	.12	.10	.07	.10	.07	.07	.09	.09	.09	.11	.09	general	78
2.3855	.11	.09	.09	.09	.09	.09	.07	.07	.08	.09	.12	as	940
2.3863	.08	.10	.10	.09	.08	.08	.10	.10	.07	.09	.12	of	3206
2.3888	.09	.09	.08	.08	.09	.10	.08	.11	.07	.08	.11	from	389
2.3892	.09	.09	.08	.08	.07	.10	.11	.09	.11	.07	.10	after	538
2.3899	.08	.11	.10	.08	.09	.08	.10	.10	.08	.08	.11	with	2315
2.3902	.09	.09	.09	.08	.09	.10	.08	.10	.07	.09	.12	early	341
2.3911	.08	.11	.10	.08	.08	.08	.10	.10	.08	.09	.11	in	2865
2.3914	.09	.11	.09	.08	.08	.09	.10	.09	.08	.08	.11	by	1408
2.3919	.09	.11	.09	.08	.08	.10	.08	.09	.08	.09	.11	course	2104
2.3936	.07	.10	.10	.08	.10	.09	.10	.10	.08	.09	.10	or	1953
2.3950	.08	.10	.10	.09	.08	.08	.09	.10	.09	.09	.10	possibly	2405
2.3955	.08	.10	.10	.09	.09	.08	.09	.09	.08	.09	.10	to	2010

Table 3. The highest 'entropy' words in CMIT. Note that these are common English words.

A	MIij	Pij	Oo	Op	Pi	Oi	Pj	Oj	Wi-Oj
0.9520	23	.96	(23 , 0)		.01 (25)	.01 (23)			vena-cava
0.9500	53	.98	(53 , 1)		.03 (103)	.02 (53)			inhalation-clv
0.9495	21	.96	(21 , 0)		.01 (22)	.01 (21)			sella-turcica
0.9492	21	.96	(21 , 0)		.01 (23)	.01 (21)			cor-pulmonale
0.9471	24	.96	(24 , 0)		.01 (46)	.01 (24)			percussion-note
0.9470	21	.96	(21 , 0)		.01 (30)	.01 (21)			lavage-catharsis
0.9450	19	.95	(19 , 0)		.01 (23)	.01 (19)			arteriosus-ductus
0.9422	27	.97	(27 , 0)		.02 (75)	.01 (27)			mg-meter
0.9384	59	.97	(58 , 1)		.03 (93)	.02 (59)			diabetes-mellitus
0.9380	33	.97	(33 , 1)		.03 (108)	.01 (33)			per-cubic
0.9321	27	.97	(27 , 0)		.03 (108)	.01 (27)			per-meter
0.9305	41	.98	(41 , 1)		.05 (150)	.01 (41)			ecg-qrs
0.9301	14	.94	(14 , 0)		.01 (23)	.00 (14)			angina-pectoris
0.9287	17	.95	(17 , 0)		.02 (60)	.01 (17)			gait-waddling
0.9279	16	.94	(16 , 0)		.02 (53)	.01 (16)			clv-vapor
0.9267	16	.94	(16 , 0)		.02 (57)	.01 (16)			occupational-vapor
0.9247	21	.96	(21 , 0)		.03 (103)	.01 (21)			inhalation-catharsis
0.9206	27	.93	(26 , 0)		.01 (33)	.01 (27)			cubic-meter
0.9191	11	.92	(11 , 0)		.00 (12)	.00 (11)			slit-lamp
0.9126	34	.94	(33 , 1)		.03 (103)	.01 (34)			inhalation-ppm
0.9126	16	.94	(16 , 0)		.03 (103)	.01 (16)			inhalation-vapor
0.9061	19	.95	(19 , 0)		.05 (150)	.01 (19)			ecg-p-r
0.9056	11	.92	(11 , 0)		.02 (56)	.00 (11)			block-bundle-branch
0.9036	29	.94	(28 , 0)		.03 (103)	.01 (29)			inhalation-manufacture
0.9035	23	.92	(22 , 0)		.02 (53)	.01 (23)			clv-percutaneous
0.9032	21	.91	(20 , 0)		.01 (31)	.01 (21)			saline-catharsis
0.8992	27	.93	(26 , 0)		.03 (103)	.01 (27)			inhalation-water
0.8974	46	.92	(43 , 0)		.02 (62)	.01 (46)			therapeutic-expected
0.8965	21	.91	(20 , 0)		.02 (53)	.01 (21)			clv-catharsis
0.8954	8	.90	(8 , 0)		.00 (14)	.00 (8)			kernig-brudzinski
0.8954	8	.90	(8 , 0)		.00 (14)	.00 (8)			leucine-aminopeptidase
0.8946	12	.93	(12 , 0)		.03 (110)	.00 (12)			fractura-comminuted
0.8912	30	.94	(29 , 1)		.05 (150)	.01 (30)			ecg-leads
0.8908	53	.93	(50 , 1)		.04 (118)	.02 (53)			air-clv
0.8904	9	.91	(9 , 0)		.02 (60)	.00 (9)			aorta-coarctation
0.8891	46	.92	(43 , 1)		.03 (89)	.01 (46)			effects-expected
0.8891	11	.92	(11 , 0)		.03 (110)	.00 (11)			nasal-rhinoscopy
0.8886	13	.93	(13 , 0)		.04 (145)	.00 (13)			jaundice-egpt
0.8881	23	.92	(22 , 0)		.03 (103)	.01 (23)			inhalation-percutaneous
0.8877	26	.96	(26 , 2)		.08 (249)	.01 (26)			cough-bronchoscopy
0.8876	29	.90	(27 , 0)		.02 (50)	.01 (29)			rhythm-gallop
0.8867	29	.90	(27 , 0)		.02 (53)	.01 (29)			clv-manufacture
0.8866	29	.97	(29 , 2)		.08 (264)	.01 (29)			anemia-normochromic
0.8866	29	.97	(29 , 2)		.08 (264)	.01 (29)			anemia-macrocytic
0.8863	12	.93	(12 , 0)		.04 (137)	.00 (12)			marrow-erythroid
0.8861	21	.91	(20 , 0)		.03 (87)	.01 (21)			gastric-catharsis
0.8835	23	.92	(22 , 0)		.04 (118)	.01 (23)			air-percutaneous
0.8833	10	.92	(10 , 0)		.03 (108)	.00 (10)			per-liter
0.8826	55	.91	(51 , 1)		.03 (96)	.02 (55)			reactions-adverse
0.8806	7	.89	(7 , 0)		.01 (26)	.00 (7)			bronchoscopy-bronchography
0.8802	34	.92	(32 , 1)		.04 (118)	.01 (34)			air-ppm
0.8793	30	.91	(28 , 0)		.03 (87)	.01 (30)			gastric-lavage
0.8768	11	.92	(11 , 0)		.05 (150)	.00 (11)			ecg-bundle-branch
0.8733	8	.90	(8 , 0)		.03 (86)	.00 (8)			murmur-holosystolic
0.8723	34	.89	(31 , 0)		.02 (53)	.01 (34)			clv-ppm

Table 4. The top of the word-pair list in decreasing order of association value (A).

A	MIij	Pij	Oo	Op	Pi	Oi	Pj	Oj	Wi-Oj
-0.1081	110	.01	(0 , 12)		.12 (381)	.03 (110)			bone-ventricular
-0.1063	91	.01	(0 , 10)		.12 (381)	.03 (91)			bone-vaginal
-0.1039	150	.01	(1 , 17)		.12 (381)	.05 (150)			bone-ecg
-0.1019	64	.02	(0 , 7)		.12 (381)	.02 (64)			bone-cervix
-0.0995	55	.02	(0 , 6)		.12 (381)	.02 (55)			bone-stricture
-0.0989	53	.02	(0 , 6)		.12 (381)	.02 (53)			bone-iris
-0.0962	51	.02	(0 , 6)		.12 (381)	.02 (51)			bone-paroxysmal
-0.0974	50	.02	(0 , 5)		.12 (381)	.02 (50)			bone-catheterization
-0.0976	50	.02	(0 , 5)		.12 (381)	.02 (50)			bone-rhythm
-0.0974	49	.02	(0 , 5)		.12 (381)	.02 (49)			bone-glaucoma
-0.0974	49	.02	(0 , 5)		.12 (381)	.02 (49)			bone-p
-0.0966	47	.02	(0 , 5)		.12 (381)	.01 (47)			bone-wave
-0.0943	93	.01	(0 , 9)		.10 (341)	.03 (93)			dyspnea-epidermis
-0.0938	41	.02	(0 , 4)		.12 (381)	.01 (41)			bone-qrs
-0.0938	170	.02	(3 , 20)		.12 (381)	.05 (170)			bone-right
-0.0932	40	.02	(0 , 4)		.12 (381)	.01 (40)			bone-sterility
-0.0926	80	.01	(0 , 8)		.10 (341)	.02 (80)			dyspnea-nerves
-0.0914	73	.01	(0 , 7)		.10 (341)	.02 (73)			dyspnea-scalp
-0.0907	36	.03	(0 , 4)		.12 (381)	.01 (36)			bone-placenta
-0.0900	35	.03	(0 , 4)		.12 (381)	.01 (35)			bone-fields
-0.0896	64	.02	(0 , 6)		.10 (341)	.02 (64)			dyspnea-urethral
-0.0893	34	.03	(0 , 4)		.12 (381)	.01 (34)			bone-corium
-0.0887	60	.02	(0 , 6)		.10 (341)	.02 (60)			dyspnea-salt
-0.0885	33	.03	(0 , 3)		.12 (381)	.01 (33)			bone-ciliary
-0.0876	32	.03	(0 , 3)		.12 (381)	.01 (32)			bone-pulmonic
-0.0875	56	.02	(0 , 5)		.10 (341)	.02 (56)			dyspnea-hyperkeratosis
-0.0872	55	.02	(0 , 5)		.10 (341)	.02 (55)			dyspnea-knee
-0.0872	65	.03	(1 , 7)		.12 (381)	.02 (65)			bone-atrial
-0.0867	64	.03	(1 , 7)		.12 (381)	.02 (64)			bone-urethral
-0.0867	97	.03	(2 , 11)		.12 (381)	.03 (97)			bone-sounds
-0.0867	31	.03	(0 , 3)		.12 (381)	.01 (31)			bone-perineum
-0.0867	31	.03	(0 , 3)		.12 (381)	.01 (31)			bone-ovary
-0.0866	53	.02	(0 , 5)		.10 (341)	.02 (53)			dyspnea-cystoscopy
-0.0866	53	.02	(0 , 5)		.10 (341)	.02 (53)			dyspnea-disk
-0.0866	53	.02	(0 , 5)		.10 (341)	.02 (53)			dyspnea-nystagmus
-0.0865	129	.03	(3 , 15)		.12 (381)	.04 (129)			bone-artery
-0.0863	52	.02	(0 , 5)		.10 (341)	.02 (52)			dyspnea-genitalia
-0.0861	95	.03	(2 , 11)		.12 (381)	.03 (95)			bone-ventricle
-0.0858	30	.03	(0 , 3)		.12 (381)	.01 (30)			bone-angiocardiography
-0.0858	62	.03	(1 , 7)		.12 (381)	.02 (62)			bone-conjunctiva
-0.0858	30	.03	(0 , 3)		.12 (381)	.01 (30)			bone-leads
-0.0858	30	.03	(0 , 3)		.12 (381)	.01 (30)			bone-exertional
-0.0855	50	.02	(0 , 5)		.10 (341)	.02 (50)			dyspnea-penis
-0.0855	50	.02	(0 , 5)		.10 (341)	.02 (50)			dyspnea-behavior
-0.0853	61	.03	(1 , 7)		.12 (381)	.02 (61)			bone-diaphragm
-0.0848	29	.03	(0 , 3)		.12 (381)	.01 (29)			bone-gallop
-0.0848	29	.03	(0 , 3)		.12 (381)	.01 (29)			bone-pupil
-0.0848	60	.03	(1 , 7)		.12 (381)	.02 (60)			bone-waves
-0.0848	29	.03	(0 , 3)		.12 (381)	.01 (29)			bone-gallbladder
-0.0848	29	.03	(0 , 3)		.12 (381)	.01 (29)			bone-dysarthria
-0.0848	29	.03	(0 , 3)		.12 (381)	.01 (29)			bone-abortion
-0.0837	58	.03	(1 , 6)		.12 (381)	.02 (58)			bone-urethra
-0.0837	28	.03	(0 , 3)		.12 (381)	.01 (28)			bone-conjunctival
-0.0837	28	.03	(0 , 3)		.12 (381)	.01 (28)			bone-field
-0.0837	28	.03	(0 , 3)		.12 (381)	.01 (28)			bone-environment

Table 5. The bottom of the word-pair list, showing the negatively correlating words.

1.2262	2	76	0	0	1	0	1	11	0	0	2	epidermis 93
1.3089	0	71	1	0	0	0	4	5	1	2	1	dermis 85
1.3902	9	59	1	0	2	1	0	4	0	0	0	papules 76
1.4672	0	37	0	0	0	0	3	4	0	0	0	acanthosis 44
1.4685	1	46	0	1	0	0	1	5	1	1	0	hyperkeratosis 56
1.6040	2	33	0	0	0	0	0	2	1	1	1	epidermal 40
1.6259	6	25	0	0	0	0	0	0	0	0	0	macules 31
1.6267	4	32	0	0	0	0	1	3	1	0	0	scaling 41
1.6619	6	50	3	1	2	0	1	1	0	0	1	scalp 73
1.7047	1	20	0	0	0	0	0	0	1	0	0	involution 22
1.7177	5	24	0	0	0	0	0	1	0	0	0	papule 32
1.7209	0	29	0	3	0	0	0	2	0	2	2	sebaceous 38
1.7246	2	19	0	0	0	0	0	0	0	0	0	horny 21
1.7307	0	18	0	0	0	0	0	1	0	0	0	keratin 19
1.7441	0	19	0	0	0	0	0	1	0	1	0	stratum 21
1.7511	10	35	1	2	1	0	1	0	0	4	0	eruption 54
1.7590	2	25	0	1	0	0	1	5	0	0	0	corium 34
1.7619	0	17	0	0	0	0	0	1	0	0	0	cornuum 18
1.7732	0	21	0	0	0	0	0	2	2	1	0	0	melanin 26
1.7819	17	98	1	3	3	3	18	34	2	0	0	pruritus 105
1.7821	3	22	0	0	0	1	0	0	1	0	0	1	pustules 28
1.8192	2	26	0	5	2	0	0	1	0	0	0	3	bullae 39
1.8388	10	24	3	0	3	0	0	0	0	0	0	0	soles 40
1.8391	0	16	0	0	0	0	0	1	0	0	0	2	scales 19
1.8395	0	16	0	0	1	0	0	0	0	0	1	0	nipple 18
1.8428	4	47	1	7	2	1	4	3	1	2	2	infiltrate 74
1.8436	0	17	0	0	0	0	1	3	0	0	0	parakeratosis 21
1.8505	10	24	3	0	2	0	1	0	0	0	0	palms 40
1.8521	1	18	0	0	1	0	1	0	0	1	0	hyperpigmentation 22
1.8560	0	16	1	1	0	0	0	0	1	0	0	cutis 19
1.8587	0	12	0	0	0	0	0	0	0	0	0	ichthyosis 12
1.9012	17	31	2	2	0	0	3	5	0	2	0	erythematous 62
1.9109	2	29	0	1	0	2	4	5	5	0	0	6	follicles 54
1.9242	0	29	0	3	1	0	2	6	0	0	5	patches 54
1.9251	0	13	0	1	0	0	0	0	0	0	2	crust 16
1.9283	0	14	0	1	0	0	0	1	0	0	1	keratosis 17
1.9337	2	20	0	0	0	0	1	1	3	2	0	3	follicular 32
1.9339	3	15	3	0	0	0	0	0	0	0	0	cheeks 21
1.9347	0	17	1	0	0	0	0	3	7	0	0	0	rete 28
1.9407	1	37	1	3	2	1	6	10	1	1	2	circumscribed 65
1.9488	0	17	0	4	0	0	2	3	0	0	1	crusting 27
1.9524	4	23	2	0	1	2	0	11	1	0	0	breast 44
1.9781	3	21	1	1	0	1	0	2	1	5	0	sweat 35
1.9765	0	10	0	0	0	0	0	0	0	0	0	subepidermal 10
1.9775	5	19	0	2	0	0	2	2	0	0	4	leaving 34
1.9787	4	37	1	4	5	0	2	5	1	6	2	plaques 57
1.9796	2	16	1	0	0	0	1	0	0	2	3	sunlight 25
1.9843	0	11	0	0	0	0	0	3	0	0	0	verrucous 14
1.9878	5	17	4	0	2	0	0	1	0	0	0	nail 29
1.9878	3	15	1	0	0	0	0	0	1	0	2	0	scaly 22
1.9883	1	16	2	0	0	0	0	0	3	0	0	0	ridges 25
1.9926	1	13	1	0	0	0	0	0	1	0	0	0	hyperkeratotic 17
1.9994	0	11	0	1	0	0	0	1	0	0	0	0	hairs 13
2.0008	3	13	0	1	0	0	1	0	0	0	0	3	eczema 21
2.0026	0	14	1	0	2	0	0	1	0	2	0	0	nevus 28
2.0032	6	20	4	0	0	0	1	1	1	5	0	0	buttocks 38

Table 6. A word list generated algorithmically which constitutes a dermatological vocabulary. The disease category 'skin' is represented by the third column.

1.4773	0	0	0	1	0	1	39	1	2	0	0	stools 52
1.5869	2	1	1	0	5	2	44	3	1	0	0	barium 59
1.5848	2	1	4	0	0	0	34	4	5	0	0	colon 50
1.6182	0	0	1	0	1	1	24	1	1	0	0	gallbladder 29
1.6338	2	0	1	0	1	2	27	1	6	0	0	duodenal 40
1.6441	1	2	0	0	3	1	18	0	9	0	0	duodenum 34
1.6627	3	0	1	1	3	4	39	21	0	0	0	peritonitis 72
1.6686	1	3	1	0	3	4	33	0	1	0	1	quadrant 47
1.6836	4	0	0	0	1	1	26	3	4	0	0	bile 39
1.6967	2	0	1	1	1	4	28	1	2	0	0	biliary 48
1.7007	0	0	1	1	2	1	33	4	9	1	0	epigastric 60
1.7381	1	0	0	0	0	0	14	0	0	0	0	gastroscopy 15
1.7445	2	0	0	0	0	7	11	0	1	0	0	urobilinogen 21
1.7659	19	0	2	0	2	0	39	6	4	6	0	constipation 78
1.7806	5	1	0	2	3	3	26	0	0	1	0	esophageal 41
1.7851	1	0	3	1	0	1	22	1	1	0	1	tooth 31
1.8025	0	0	0	0	0	0	11	0	2	0	0	jejunum 13
1.8077	0	3	0	0	0	0	6	11	0	1	0	0	pulp 21
1.8145	1	0	1	0	0	0	0	14	1	0	0	0	colonic 17
1.8187	0	0	1	0	0	0	0	13	1	0	0	0	emesa 15
1.8188	3	0	0	0	0	0	0	13	0	0	0	0	lap 16
1.8410	2	0	1	0	0	0	0	13	0	0	0	0	pyloric 16
1.8424	3	1	0	2	0	0	0	15	0	0	0	0	submaxillary 21
1.8687	7	0	0	1	1	0	21	2	3	1	0	bilirubin 42
1.8692	12	5	0	2	0	0	25	4	1	1	0	feces 50
1.8785	0	0	0	0	0	0	0	10	0	0	0	0	periportal 10
1.8741	2	1	0	1	0	0	0	13	1	3	0	0	meal 21
1.8757	0	0	0	1	0	0	0	11	1	2	0	0	cecum 15
1.8842	3	1	0	1	1	1	1	21	3	0	4	0	stool 35
1.8897	9	4	1	2	2	2	7	30	0	3	2	0	cirrhosis 60
1.8975	4	2	0	0	3	1	16	0	0	0	0	mesenteric 26
1.8987	2	1	0	0	1	1	15	1	1	0	0	peristalsis 22
1.8991	1	0	0	0	1	0	0	11	0	0	0	0	sept 13
1.9080	0	0	1	0	0	0	0	11	1	0	0	0	proctoscopy 13
1.9060	6	1	5	0	5	5	33	4	2	3	0	intestine 64
1.9172	0	0	0	0	0	0	0	9	0	0	0	0	cholangitis 9
1.9172	0	0	0	0	0	0	0	9	0	0	0	0	cholecystography 9
1.9172	0	0	0	0	0	0	0	9	0	0	0	0	esophagoscopy 9
1.9224	4	5	1	0	0	0	0	15	0	0	1	0	anal 26
1.9238	1	0	0	0	3	4	10	0	1	0	0	varices 19
1.9634	0	0	0	0	1	1	9	0	0	0	0	intrahepatic 11
1.9728	1	0	0	0	0	2	6	0	2	0	0	gastrectomy 11
1.9736	0	0	0	1	0	0	9	1	1	0	0	intussusception 12
1.9773	0	2	0	0	0	0	10	6	2	0	0	loops 20
1.9775	1	1	0	1	2	5	14	1	1	1	1	portal 28
1.9812	0	1	0	0	0	0	0	0	1	0	0	jejunal 16
1.9815	0	0	0	0	0	1	5	0	2	0	0	aminopeptidase 8
1.9841	3	0	0	0	0	0	0	9	0	0	0	0	thymol 12
1.9872	0	1	2	1	0	0	1	11	1	0	0	0	sigmoid 16
1.9888	2	2	0	3	0	0	0	13	4	0	0	0	submucosa 24
1.9890	1	1	0	0	2	0	0	11	1	1	0	0	ileum 17
1.9933	3	0	0	0	0	3	0	8	0	0	1	0	achlorhydria 15
2.0083	1	0	0	0	0	1	0	10	0	1	2	1	parotid 16
2.0093	0	0	0	2	0	0	0	11	3	0	1	2	polyps 19
2.0099	0	0	0	0	0	0	0	3	0	2	0	0	subtotal 5
2.0109	2	2	0	1	0	0	0	10	0	0	0	0	colitis 15

Table 7. A word list generated algorithmically which constitutes a vocabulary of gastroenterology. The eighth column represents the disease category 'digestive system'.