the passive because it happens to be a handy means of indicating the performance of an act, particularly when the subject of the action is too unimportant or too obvious to mention (*e.g.*, "The model is developed...," "Accurate rate constants can be determined...," etc.). The passive connotation is also present in certain elliptical constructions of the type N + Ven (past participle): *e.g.*, "the alloys employed...," "radical compressive stress caused...," etc., which imply a relative clause with a passive verb ("alloys which are employed," "radical stress which is caused," etc.). The use of N + Ven is only matched in popularity by the use of N + Ving (present participle) constructions, as in: "formula describing...," "change occurring...," "alloys contg...," etc.

One of the devices profitably exploited for sentence modification in *CA* is the prepositional phrase. A number of important facts, conditions, situations, etc. are expressed concerning the action indicated by the verb of the sentence (*e.g.*, "One, or possibly two, unidentified metastable carbides may be precipitated from supersaturated Si Ferrite at 1100° F. and below." "The martensitic types transform to martensite on cooling to room temp.," etc.). The proportion of other sentence adverbials is very low.

2. THE SUBJECT

The subject types in the corpus seem to follow certain specific patterns in their structure. The majority of them are expansions of one kind or another, and are very flexible. Possibilities made use of include: single noun, row of nouns joined by conjunction, nominal compound or phrase, noun followed by prepositional phrase as post-modifier, noun followed by relative clause, gerund, etc. Of these different categories, the nominal phrases and compounds acting as subjects deserve some illustration. They are utilized by the abstractor to the fullest extent because of their compactness and economy. Some examples are (a) *Compounds*: "source unit" (unit which is the source), "constitution diagram" (diagram of constitution), "solute segregation" (segregation of solute), "quartz spectrograph" (spectrograph of quartz), "blast furnace smelting" (smelting in blast furnace), "water quenching" (quenching by water), "kitchen utensils" (utensils for kitchen), "corrosion resistance" (resistance to corrosion), "anti-tarnish protection" (protection against tarnishing), "straight line relation" (relation like that of a straight line), etc.; (8) *Phrases*: "segregation of carbide" (derived noun + prepositional object), "composition of steel" (property of material), "pelleting of ore-concentrates" (gerund + prepositional object), etc.

3. THE COMPLEMENT

For purposes of our study we stretch the use of the term "complement" to include all linguistic material that follows the verb in a regular-order sentence structure. An analysis of the corpus has revealed that the following types of complements commonly occur after the verbal: zero (no complement), noun (acting as complement of linking verb), adjective (acting as complement of linking verb), direct object, and prepositional phrase. Of these, the prepositional phrase is the most frequent and the most versatile.

Before closing, one or two uses of this study may be mentioned. Any kind of processing of natural language text requires a thorough knowledge of its structural as well as stylistic features. In information retrieval the methodology for structuring and querying the file is greatly dependent upon the stylistic peculiarities of the text. Further, a comparison of *CA* with other technical abstracts can be attempted on the basis of some of the data presented here to see how a maximum of information can be conveyed in a minimum of space.

A Study of Coverage in *Chemical Abstracts* of the Literature on C-Reactive Protein, a Biomedical Borderline Area*

RICHARD F. RILEY, ROBERT LEWIS, HARVEY S. FREY, and Y. HOKAMA
Center for the Health Sciences, University of California, Los Angeles, California

Received September 23, 1965

This study is an outgrowth of two activities of the senior author (R. F. R.): as a scientist involved in biomedical research and as a part-time abstractor for the Chemical Abstracts Service (CAS). Assignments to abstractors come from CAS either as references to specific papers or as a responsibility for covering all suitable material in a specific journal. Obviously, the selection of a paper for abstracting requires the decision to be made by someone that that particular paper contains material of chemical interest (*1*). Personal experience has shown that making

*The bibliographic file studied in preparation of this report was accumulated during the course of other research which was supported in part by Research Grant 5 RO 1-HE 06512 from the National Heart Institute.

The title study shows that coverage for the period 1930–1961 was consistent from year to year, and bias toward publication in any of the major languages was absent. From an analysis of information contained in abstracts of articles with that contained in unabstracted papers, it is concluded that relatively little significant new chemical information was missed. Among other factors which are discussed, the quality of a particular study may predicate for or against the likelihood of its being abstracted.

such decisions can be difficult in the case of journal assignments in borderline medical areas.

As scientists engaged in laboratory studies in a biomedical area, we have assembled an extensive annotated bibliography relevant to our area of research for the usual reasons one does this. The bibliography, described in more detail later, has been annotated by addition of extensive reading notes and abstracts from *CA*. From a casual comparison of articles for which abstracts had appeared in *CA* with those which had not been abstracted, a subjective impression was obtained that abstractors, for reasons which are not apparent, perhaps were not entirely consistent in deciding what is "chemical" in the borderline area in which we work. The analysis which follows examines the factors which may have operated in the selection of certain borderline material for inclusion in *CA* and the consistency with which selections have been made. The study is restricted to *CA* for the period 1930 through 1961 for reasons noted later.

THE BIBLIOGRAPHIC SAMPLE

The bibliographic material from which our sample was drawn consists of about 1200 references from 1930 to December 1965, directly or indirectly related to C-reactive protein. This protein is not present normally in serum or serous exudates but appears therein as a consequence of a variety of pathological processes. It is therefore the subject of many clinical studies and, because of its possible physiological importance, is the subject of a smaller number of studies which have appeared in journals devoted to the basic sciences. The bibliography was assembled from entries found in *CA*, *Excerpta Medica*, *Current List of Medical Literature*, *Index Medicus*, especially bibliographies to review articles on the subject, in the articles themselves, and from other miscellaneous sources. The bibliography is thought to be representative of the available world literature on the subject and to be reasonably complete, since continuing search is not turning up significant numbers of new references to old literature.

For this study it was necessary to eliminate a number of the available references from consideration. These included editorial and anonymous material, abstracts of papers presented at scientific meetings, and some other types of published and unpublished information which are not usually considered suitable for abstracting by CAS. Further, all references appearing after 1961 were eliminated in order to provide a sample for which a reasonable time had elapsed since publication for appearance of an abstract in *CA*. With these deletions 679 references remained which might have conceivably provided abstractable material.

PRELIMINARY ANALYSIS OF THE SAMPLE

Prior to analysis each reference was transferred to an individual file card on which was recorded authors' names, title of the article, journal, volume, page, year, and language of publication. As analysis progressed, additional information was recorded on the cards for eventual transfer to IBM cards for computer handling.

As to be expected, the 679 references are to articles in journals and fall into the following three categories: *category A*, articles from journals not scanned for suitable material by CAS or their abstractors; *category B*, articles from journals scanned by CAS or their abstractors for which no abstract had appeared; and *category C*, articles for which abstracts had appeared in *CA*. References in category C were identified by independent searches by two individuals of the author indexes of *CA*. Abstracts of work published prior to 1959 were sought in issues of *CA* which appeared during the succeeding five-year period after publication of the paper. Abstracts to more recent citations were sought in indexes which appeared for at least the succeeding two years after publication of the work. There were 327 references in category C. Distinction between journal articles falling in categories A and B was made by reference to the CAS lists of periodicals scanned and, in instances where some question existed, by correspondence with CAS. Totals of 66 and 296 references fell in categories A and B, respectively. The total sample was scattered through 330 different journals and appeared in 19 languages. Further characterization of the sample by year of publication, language, and journal category is given in Tables I and II, and is discussed later.

SUBJECT ANALYSIS OF THE SAMPLE

We wished to inquire what kind and how much information of a type editorially admissable by CAS occurs in references in categories A and B and how much of such information is in articles in category C but not in the published abstracts. Therefore it was necessary to obtain first a general idea of the types of information contained in CAS abstracts of references in category C.

Study of the subject content of CAS abstracts of journal articles in category C showed that their information content could be described reasonably well by one or more of the following general descriptions:

1. Statistical data on the frequency of appearance of C-reactive protein in the patient's serum of other body fluids (single determinations).
2. Correspondence of the appearance of C-reactive protein in serum in disease in comparison with other chemical and physiological indices of disease (single determinations).
3. Sequential changes in this and other indices of pathology during progression of or recovery from disease.
4. Effects of pharmacological agents on this and other indices of disease.
5. Experimental studies on this substance in animals and effects of C-reactive protein on physiological systems.
6. Biochemical and biophysical studies of this protein, except 7.
7. Immunoelectrophoretic studies on this and other serum proteins.
8. Statements of utility of this and other clinical indices of pathology.
9. Studies of methods for this and other chemical indices of pathology.
10. Behavior of this protein with respect to passage through tissue barriers.
11. Article stated to be a review.
12. Abstracted "by title only".

The type of information contained in abstracts of references in category C matching these descriptions was noted on the corresponding file cards for each. All 679 journal articles were then examined for the presence of any of the first 11 sorts of admissable information contained therein, and this information was also added to each file card. Since there appeared to be great variation in the amount of correlative information with respect to C-reactive protein and other chemical and physiological indices studied (type 2 above), the extent of the latter was scored separately.

Finally, each study was judged at the time of examination to be an excellent, good, fair, or poor piece of work and graded from 1 to 4, respectively. A rating of poor was accorded those studies carried out on inadequate numbers of poorly or uncharacterized samples and in which the data were presented in such a manner as to preclude evaluation. An excellent score was assigned studies clearly showing evidence of well-thought-out experimental design or utilizing a new point of departure, carried out by reputable methods using valid numbers of properly characterized samples and, where appropriate, analyzed statistically. Assignments of good or fair were made to citations intermediate in these regards.

These data were transferred finally to IBM cards and processed by appropriate programs by the Health Sciences Computing Facility, UCLA. The statistic employed was the chi-squared test of independence.

RESULTS

Table I gives the distribution of the references in the sample by year of publication for the three categories. About 10% of the references appeared in journals which were not being scanned by CAS at the time they appeared. Roughly one-half of the studies which appeared in journals being scanned for appropriate material were abstracted. The fraction of the articles in journals being scanned which were abstracted did not differ significantly by year ($P > 0.10$).

Table II gives the distribution of the articles in the sample by the language in which they were published and their reference category. Here again, the fraction of articles being scanned which were abstracted did not differ significantly with respect to the language in which the work was published ($P > 0.10$).

While grading of the quality of the reported work was admittedly subjective, it was felt that quality might operate for selection or rejection of material for abstracting. Table III lists the distribution of the sample by quality for the three reference categories. To the extent our subjective impression of quality is valid, the percentage of references abstracted from journals scanned was highly correlated with quality ($\chi^2 = 16.03$, $df = 3$, $P < 0.005$) in favor of better reports, and the same was true for the total sample.

Table I. Distribution of the Sample by Year and Reference Category

Year of publication	Category A	B	C	Total	% of references abstracted of those scanned[a]
1961	7	26	40	73	60.5
1960	19	36	42	97	55.2
1959	12	36	54	102	60.0
1958	14	53	53	120	50.0
1957	8	50	58	116	53.7
1956	1	28	29	58	50.9
1955	3	27	12	42	30.8
1930–1954	2	30	39	71	56.5
Total	66	286	327	679	
% total sample	9.7	42.1	48.2	100	

[a] C/(B + C) × 100.

Table II. Distribution of the Sample by Language and Reference Category

Language	Category A	B	C	% of references abstracted of those scanned[a]
English	8	118	142	54.5
Italian	38	83	110	57.0
German	1	18	22	55.0
French	6	15	15	50.0
Russian	..	15	17	53.0
Others	13	37	21	36.2
Total	66	286	327	53.4

[a] C/(B + C) × 100.

Table III. Distribution of the Sample by Subjective Impression of Quality and Reference Category

Quality	Value assigned	Category A	B	C	% of references abstracted of those scanned[a]
Excellent	1	0	19	37	66.0
Good	2	8	102	138	57.5
Fair	3	24	96	110	53.4
Poor	4	34	69	42	37.8
Mean values		3.39	2.75	2.48	

[a] C/(B + C) × 100.

Chi-square tests of independence were made for each of the 11 general types of subject matter found in abstracts of references in category C with those found in references in category B. The results indicate that reported sequential changes, experimental work in animals, obvious biochemical data, immunoelectrophoretic studies, and studies of methods (types 3, 5, 6, 7, and 9 above) predicated for inclusion in *CA*. Statistical data and statements of utility (types 1 and 8 above) were more prevelant in articles which had not been abstracted than in those which were. Information on correlations of the presence of C-reactive protein in serum with results of other tests, drug effects on its behavior, its passage through tissue barriers, and reviews (types 2, 4, 10, and 11 above) were indifferent indicia. These data are included in Table IV. It is also evident from the data in Table IV that an appreciable amount of material of a type acceptable to *CA* has appeared in scanned but unabstracted articles. In striking contrast, a comparison of the types of information found in articles in category C with those found in their corresponding abstracts showed that 94% of the admissable information in the article was noted in the abstract. Of those admissable types of information predicating for inclusion in *CA* (types 3, 5, 6, 7, and 9 above) which were found in the article, 98% were noted in the abstracts.

From the ratios given at the bottom of Table IV it would appear on the average that more types of information were present in the order: category A < category B < category C. That is, references in category C tend to be from more extensive studies, and this is in line with our subjective impression of quality.

As noted earlier there was considerable variability in the number of other tests which were carried out and correlated with the presence or absence of C-reactive protein in the samples analyzed. For example, one study might entail a comparison of C-reactive protein with transaminase levels in serum, while another concerned with the same comparison might include other "bits" of data of chemical and/or physiological significance—*e.g.*, sedimentation rate, concentration of serum protein fractions, nonspecific colloid lability reactions, etc. Occurrence of such other "bits" of chemical and physiological types of information were tallied separately for all references containing type 2 information. The data are tabulated in Table V.

Table IV. Distribution within the Sample of the Frequency with Which Various Types of Information Appeared Compared by Reference Category

Type of information[a]	Category A	B	C	*P*
Statistical data(1)	35	133	111	<0.005
Test correlations(2)	26	131	154	N.S.[c]
Sequential changes(3)	3	36	61	0.05
Drug effects(4)	6	25	39	N.S.
Animal studies(5)	0	11	32	0.005
Biochemistry(6)	0	5	39	<0.005
Immunoelectrophoresis(7)	1	10	22	0.1
Utility(8)	8	61	49	0.05
Methods(9)	2	16	36	0.025
Tissue barriers(10)	5	13	16	N.S.
Reviews(11)	4	22	16	N.S.
Title only(12)	0	0	4	Inadequate sample
Density of types of information per citation[d]	1.36	1.62	1.77	

[a] Described in detail under "Subject Analysis of the Sample." [b] Values based on information in published abstracts. [c] N.S. = not significant. [d] Calculated as the sum of the column divided by the number of references in each reference category.

Table V. Distribution by Reference Category of Numbers of Other Ancillary Tests of Chemical and Physiological Nature in Articles Containing Type 2 Information

	Category A	B	C
Extra "bits" in article or abstract			
Chemical	12	82	111
Physiological	24	99	78
Extra "bits" in article not indicated in abstract			
Chemical	..	...	30
Physiological	..	...	44
Total number of references containing type 2 information	26	131	154

DISCUSSION

The data shown in Tables I and II indicate that abstracting performance has been consistent over the past few years and that bias toward publications in any of the major languages is absent. This is consistent with the long-standing aim of CAS to cover completely the world's chemical literature.

The large fraction of the sample judged to be of fair or poor quality is thought to be a reflection of the nature of the sample rather than the severity of the grader. The sample is weighted heavily with poor clinical reports which have appeared in many minor medical journals, especially of Italian origin. It seems quite likely that abstractors, perhaps subconsciously, many times are inclined to dismiss trivial papers as unsuitable for abstracting than those of better quality, even though trivial papers may contain admissable information.

The indication (Table IV) that biochemical data, experimental work in animals, method studies, and immunoelectrophoretic studies predicated for abstracting is not surprising. Sequential changes probably predicated for inclusion since they were described more commonly in better papers than in poorer ones. References containing material of these general descriptions had mean quality values ranging from 1.77 (biochemical data) to 2.14 (sequential changes). It is not surprising that statistical data were frequently ignored as this is stated CAS policy (*1*). Nevertheless, this was essentially the total content of a number of abstracts (*e.g., Chem. Abstr.*, **52**, 6583*h*; **55**, 2764*f*; **50**, 2817*i*). Papers containing such data and no other admissable information were the worst of the sample with a mean quality value of 3.38.

Statements of utility were recorded only where they were the prominent conclusion of the study. In many instances such statements were based on "clinical impression" and possibly for this reason were usually disregarded by abstractors. It was somewhat surprising that

drug effects on the C-reactive protein response and correlations of its presence with other tests were indifferent indicia. While neither of these topics was analyzed in finer detail, several impressions regarding them were formed in the course of the study. Probably drug effects commonly were not noted because of the clinical nature of many of the reports. Where drug effects were studied in more experimental situations either in humans or in animals, such studies usually were abstracted. Correlations of the presence of C-reactive protein with other test results often were not noted, probably because comparisons were frequently to nonchemical measures such as erythrocyte sedimentation rate, electrocardiographic changes, etc. A measure of this is included in Table V. We also have the impression that considerable variation exists in reporting the many relative nonspecific reactions given by serum such as the thymol turbidity reaction, Weltman reaction, etc., which account largely for the missed "chemistry" shown in Table V. Reviews which were not accorded an abstract were those of a predominantly clinical nature with relatively little on the subject, trivial reviews of a few papers, or reviews with inadequate bibliographic documentation.

The most interesting finding is the relatively high density of admissable material in references in category B in comparison with the small percentage of admissable information in references in category C which was missed in their abstracts. In an attempt to gain some insight into the reasons for this, references in categories B and C were matched by topic content for reexamination. For example, references in category B containing type 1 information were sorted for comparison with those in category C containing type 1 information, and type 2 for comparison against type 2, and type 1 plus 2 against 1 plus 2, etc. for the two journal categories.

Reexamination failed to reveal, in a number of instances, why one of such paired articles had been abstracted and the other not. The following paired examples are representative of such references in categories B and C:

Type 7 and 11 information;

Scheidegger, J. J. (Immunoelectrophoretic analysis of biological fluids), *Bull. soc. chim. biol.*, **39**, Suppl. 1, 45–63 (1957); not abstracted by CAS.

Scheiffarth, F., and H. Gotz (Significance of immunoelectrophoresis in differentiation of pathological sera), *Intern. Arch. Allergy Applied Immunol.*, **16**, 61–92 (1960); *Chem. Abstr.*, **54**, 13361*b* (1960).

Type 9 and 11 information:

Wood, H. F., and M. McCarthy (Laboratory aids in the diagnosis of rheumatic fever and in evaluation of disease activity), *Am. J. Med.*, **17**, 768–74 (1954); not abstracted by CAS.

Wunderly, C. (Clinical chemistry of proteins in normal and pathologically altered blood serum), *Acta Haematol.*, **20**, 9–26 (1958); *Chem. Abstr.*, **52**, 18779*g* (1958).

Type 1, 2, and 8 information:

Kasalitsa, C. L. (C-reactive protein in differential diagnosis of myocardial infarction and stenocardia), *Sovet. Med.*, **24**, (1) 63–65 (1960); not abstracted by CAS.

Carbajal, B., *et al.* (C-reactive protein in myocardial infarction and acute coronary insufficiency), *Rev. asoc. bioquim. Arg.*, **23**, 166–173 (1958); *Chem. Abstr.*, **53**, 12449*i* (1959).

Type 2 and 3 information:

Boltax, A. J., and E. E. Fischel (Serolgic tests for inflammation. Serum complement, C-reactive protein, and erythrocyte sedimentation rate in myocardial infarction), *Am. J. Med.*, **20**, 418–427 (1956); not abstracted by CAS.

Zemskov, V. M. (Dynamics of C-reactive protein, erythrocyte sedimentation rate, and leucocytes in myocardial infarction), *Lab. Delo*, **7**, (3) 20–22 (1961); *Chem. Abstr.*, **55**, 21333*a* (1961).

The most common distinction between references in category B and those in C, when paired by subject content, was in the title of the article. Those references which were abstracted frequently had more chemical sounding titles, while those which were not usually had titles with a more clinical flavor. An example may be noted in the second pair of citations given above. It seems likely that a considerable fraction of the admissable material in references in category B was missed as a result of decision making based on title alone.

Probably the most important factor to the user of *CA* is the question of how much new, significant chemical information is missed. To gain some impression on this point, studies considered to be good or excellent belonging to categories A and B were segregated into groups containing information predicating for inclusion in *CA* (types 5, 6, 7, and 9). This rather arbitrary division was chosen to yield the majority of items of significant, new, relatively certain chemical interest and a minority of those with more strictly clinical interest. The yield of unabstracted new, significant material consisted of three studies possessing some degree of originality on methods of determining C-reactive protein (all using immunological technics); two studies containing new information on its electrophoretic behavior, of which one also included new data on an isolation method; three reports on experimental studies in animals (largely immunological); and one on the relationship of the presence of C-reactive protein in serum to the positivity of several chemical tests purported to detect cancer. Only one of these nine was a reference in category A. Clearly nothing of more than trivial chemical interest had been missed. It should be noted that these figures are not in accord with those in Table IV because of the selection of better references for consideration and because of the excessively repetitious nature of the clinical work reported in the sample.

Finally, it appears to the authors that no loss of significant information to the user of *CA* would occur if abstractors ceased abstracting articles containing only type 1 and/or 2 information. Determinations of a great many of the large number of clinical indices of disease are essentially worthless in the majority of contexts unless made repeatedly. Abstracts reporting statistical results of these types accounted for slightly more than 30% of the abstracted papers in the sample.

LITERATURE CITED

(1) "Direction for Abstractors," The Chemical Abstracts Service of the American Chemical Society, Columbus, Ohio, Oct. 1, 1964.