Agreement between Medline searches using the Medline-CD-Rom and Internet PubMed, BioMedNet, Medscape and Gateway search-engines

ROSA ANGELA CARO-ROJAS and
JAVIER H. ESLAVA-SCHMALBACH

Recibido 6 Diciembre 2004/Enviado para Modificación 25 Abril 2005/Aceptado 25 Mayo 2005

ABSTRACT

Objective To compare the information obtained from the Medline database using Internet commercial search engines with that obtained from a compact disc (Medline-CD).

Methods An agreement study was carried out based on 101 clinical scenarios provided by specialists in internal medicine, pharmacy, gynaecology-obstetrics, surgery and paediatrics. 175 search strategies were employed using the connector *AND* plus text within quotation marks. The search was limited to 1991-1999. Internet search-engines were selected by common criteria. Identical search strategies were independently applied to and masked from Internet search engines, as well as the Medline-CD.

Results 3,488 articles were obtained using 129 search strategies. Agreement with the Medline-CD was 54% for PubMed, 57% for Gateway, 54% for Medscape and 65 % for BioMedNet. The highest agreement rate for a given speciality (paediatrics) was 78,1 % for BioMedNet, having greater -/- than +/+ agreement.

Conclusions Even though free access to Medline has encouraged the boom and growth of evidence-based medicine, these results must be considered within the context of which search engine was selected for doing the searches. The internet search engines studied showed a poor agreement with the Medline-CD, the rate of agreement differing according to speciality, thus significantly affecting searches and their reproducibility. Software designed for conducting Medline database searches, including the Medline-CD, must be standardised and validated.

Key Words: Medline, Databases, PubMed, reproducibility of results, information storage and retrieval, evidence-based medicine (*source: MeSH, NLM)*.

RESUMEN

Concordancia de las búsquedas de Medline utilizando el Medline en disco compacto y los motores de búsqueda de Pubmed, Biomednet, Medscape y Gateway

Objetivo Comparar la información obtenida de la base de datos Medline a través de los motores de búsqueda de Internet comerciales, con aquella obtenida de la base de datos en disco compacto (Medline-CD).

Métodos Estudio de concordancia en los resultados de las búsquedas, hechas a partir de 101 escenarios clínicos sugeridos por especialistas de medicina interna, farmacia, ginecología y obstetricia, cirugía y pediatría. Se realizaron 175 estrategias de búsqueda utilizando el conector *AND* y texto entrecomillado. La búsqueda se limitó al periodo 1991-1999. Los motores de búsqueda de Internet se seleccionaron a partir de criterios comunes. Las estrategias de búsqueda fueron idénticas y se aplicaron de manera independiente y enmascarada tanto en los motores de búsqueda de Internet como en Medline-CD.

Resultados Se obtuvieron 3 488 artículos utilizando 129 estrategias de búsqueda. La concordancia con Medline-CD fue del 54 % PubMed, 57 % Gateway, 54 % Medscape and 65 % BioMedNet. El mayor nivel de acuerdo por especialidad (pediatría) fue de 78,1 % para BioMedNet, teniendo mayor acuerdo -/- que +/+.

Conclusiones Aunque el acceso libre a Medline ha potencializado el desarrollo de la Medicina Basada en Evidencia, los resultados de las búsquedas deben ser considerados a la luz de cuáles motores de búsqueda se utilizaron. Los diferentes motores de búsqueda tuvieron una pobre concordancia con Medline-CD, siendo diferencial el nivel de acuerdo, por especialidad. Esto afecta la reproducibilidad de las estrategias de búsqueda. Todo el software que se utilice para llevar a cabo búsquedas en la base de datos Medline, incluido el de Medline-CD, debe ser estandarizado y validado.

Palabras Clave: Medline, PubMed, Bases de Datos, Reproducibilidad de resultados, recuperación, información, medicina basada en evidencia (*fuente: Decs, BIREME*).

The boom in evidence-based medicine (EBM) has meant that people working in the field of health can review the literature looking for better evidence to sustain their decisions. Medline is one of the most frequently used sources of information today (1-3), providing the source of reference for those articles which may be used by doctors when taking clinical action. Internet is the main tool for gaining access to Medline as it is fast and economic; it also allows wide scale searches to be made from 1966 to date (when using some search engines). Advanced searches can be made with *AND, OR, NOT* connectors (4) and abstracts may be obtained for articles of interest found during a particular search.

The number of servers enabling Medline searches has grown; however, each has different restrictions regarding their use in terms of the range of years available for a search, advanced search availability, etc., making each one heterogeneous. There were 37 referenced servers enabling access to Medline in 1998 (5). However, the software enabling searches to be made has not been developed as part of a standard process applicable to all servers; it has resulted from local efforts regarding each server. It has thus become very important for those using the literature and those software developers who make the searches to know whether information provided by servers enabling searches concerning Medline data-bases has sufficient agreement with results obtained from searches using original CDs. The quality of information obtained from different general access data sources in the network has been frequently evaluated (1,6); such evaluation has provided poor results in terms of the quality of such information being provided from several sources.

This study evaluated agreement between information obtained by using free Internet search engines operating from Medline databases respecting that information obtained in searches using a Medline compact disc (Medline-CD) to find out which of them had greater agreement according to search specialisation. We did not find another agreement evaluation between information provided by search engines facilitating such searches in Medline and Medline-CD. Software developed for such end, included Medline-CD's, (in different places) has still not been standardised, being dependant on each server's local development and maintenance.

METHODOLOGY

Agreement was studied between NCBI PubMed (7) (PubMed), NLM Gateway-Medline/PubMed (8) (Gateway), Medscape-Medline Search (9) (Medscape), BioMedNet (10) (BioMedNet), with Sylver Platter Int. Medline (11) (Medline-CD) suppliers. Inclusion criteria for these search engines were: having free access via Internet; having Medline database available from 1966 to October 2001; being widely used; allowing advanced searches to be made; and allowing searches to be made through operators using connectors and limiters. The period spanning 1991 and 1999 was imposed as a limit restricting searches, as this coincided with the Internet boom and with some editorially uniform manuscript standardisation processes. Including search engines having databases available from 1966 guaranteed that the search engines were as similar as possible.

Experts in each area (internal medicine, gynaecology-obstetrics, surgery, paediatrics and pharmacy) constructed 20 relevant clinical questions which were transformed into answerable questions in terms of EBM (12). Each question generated at least one strategy in terms of its search which was identically reproduced in each chosen server and the reference CD. These strategies included the *AND* operator and using text within inverted commas when required. 101 questions were included. The doctors drew up two different search strategies using 74 questions and one search strategy having 27 questions, leading to 175 search strategies being obtained (similarly used in each search engine and the Medline-CD).

Sample size was estimated, taking into account a 5 % alpha error, 20 % beta error, 85 % probability of correct classification and 0.1 difference between expected and observed agreement, implying a minimum 1 202 articles which had to be found by the search engines and the Medline-CD (13).

After a search strategy had been decided upon, each search engine was accessed and a search ordered of Medline (where necessary). Advanced searches were then ordered, limited to 1991-1999. No limits were imposed in terms of author, type of publication, language or other parameter. The query box or search box initially offered by the search engine was used; no other type of search tool was used (i.e. those offering greater search exactitude). This does not mean that search engines were not exploited to their maximum potential, but rather that they were tested according to general searches which might be made when using a CD (no advanced search tools being available) by a non-expert user. This strategy, additionally, increased the potential comparability between different search engines. All searches were done by one of the authors.

Search results were then ordered according to specialisation, question number and search strategy. Search strategy inclusion criteria were those suggested by the doctors consulted. Exclusion criteria stated that a search would generate more than 100 articles or when search engine results produced 20 or more times the number of articles found on the Medline-CD.

All articles produced by any search involved in the study were filed so that there was no probability of observer bias (demanding blind sampling).

Each article found by whichever search engine (3 488 articles) was used as a unit of measurement and agreement between search engines and Medline-CD. Each article was classified as being found or not found by a particular search engine; this information was then included in a database. Other variables which were included were search strategy code, specialisation and Medline-CD strategy result. Those articles repeatedly found by another strategy were included as being new ones, with the respective results by search engines for this different strategy. Each article was analysed from this database to verify whether it had been identified by the Medline-CD and by each search engine; agreement could then be established. 2*2 Tables were then produced and agreement measured using unweighted Kappa values for dichotomous variables. A +/+ and -/- agreement was measured. A +/+ agreement means a concordance between the search engine tested and Medline-CD to find the same articles and a -/- agreement means a concordance between both, in not to find articles with each one of the search strategies. Each clinical area included was also tested. Stata software (v. 6.0) was used for the above analysis.

RESULTS

46 out of the 175 search strategies initially suggested by the doctors were excluded (Figure 1). A search strategy for BioMedNet and another for Medscape were excluded as they produced more than 100 articles (Figure 1).

BioMedNet and Gateway had the greatest general agreement with the Medline-CD (Table 1).

Table 1. General agreement amongst evaluated search engines *cf* Medline-CD +/+ and -/- agreement

Source	Observed agreement (%)	+/+ agreement	-/- agreement	Kappa	p value
PubMed	54,1	30,3	23,8	0,1622	0,0000
Gateway	57,9	32,3	25,7	0,2327	0,0000
Medscape	54,8	32,3	22,5	0,1918	0,0000
BioMedNet	65,8	27.0	38,8	0,3188	0,0000

Tables 2, 3, 4 and 5 show agreement in different specialisations observed for each search engine. Search engines showing the highest specialisation agreement values were as follows: 63,1 % for internal medicine using BioMedNet; 55,4 % for pharmacy using BioMedNet; 75,9 % for gynaecology-obstetrics using Medscape; 66,4 % for surgery using Medscape; and 78 % for paediatrics using BioMedNet.

Table 2. Agreement observed for PubMed (*cf* the Medline-CD) for each of the study's specialisations

PubMed	Observed agreement (%)	+/+ agreement	-/- agreement	Kappa	p value
Internal medicine	45,4	21,1	24,3	0.006	0.4126
Pharmacy	44,7	24,0	20,7	0.097	0.0003
Gynaecology-obstetrics	45,2	28,3	16,9	-0.076	0.9707
Surgery	57,6	45,4	12,2	0.164	0.0000
Paediatrics	67,1	30,4	36,8	0.390	0.0000

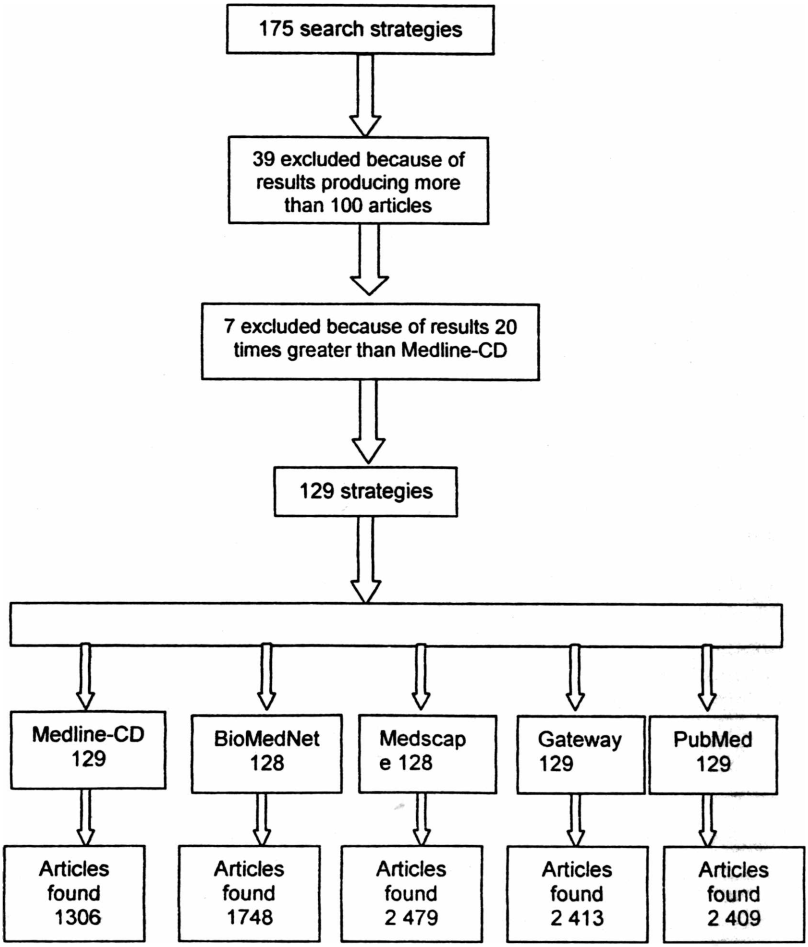
Table 3. Agreement observed for Gateway (*cf* the Medline-CD) for each of the study's specialisations

Gateway	Observed agreement (%)	+/+ agreement	-/- agreement	Kappa	P value
Internal medicine	60,0	21,3	38,8	0.2007	0.0000
Pharmacy	45,1	24,0	21,1	0.1008	0.0002
Gynaecology-obstetrics	59,1	41,4	17,8	0.2101	0.0000
Surgery	53,0	46,4	6,7	0.0787	0.0001
Paediatrics	65,7	29,4	36,4	0.3606	0.0000

Table 4. Agreement observed for Medscape (*cf* the Medline-CD) for each of the study's specialisations

Medscape	Observed agreement (%)	+/+ agreement	-/- agreement	Kappa	P value
Internal medicine	38,3	16,4	22,1	-0.0783	0.9965
Pharmacy	51,1	23,5	27,7	0.1404	0.0000
Gynaecology-obstetrics	75,9	44,7	31,3	0.5281	0.0000
Surgery	66,4	42,1	21,4	0.3361	0.0000
Paediatrics	47,6	31,0	16,6	0.1413	0.0000

Figure 1. Search strategy inclusion sequence



Gateway had the greatest +/+ agreement, whilst BioMedNet had the best -/- agreement. BioMedNet had most +/+ agreement for internal medicine, Gateway having most -/-agreement. PubMed and Gateway had greater +/+ agreement for pharmacy, BioMedNet having the greatest -/- agreement. Medscape had the greatest +/+ agreement for gynaecology-obstetrics, BioMedNet having the greatest -/- agreement. Gateway had the greatest +/+ agreement for surgery, BioMedNet having the greatest -/- agreement. Medscape had the greatest +/+ agreement for paediatrics, BioMedNet having the greatest -/- agreement.

Table 5. Agreement observed for BioMedNet (*cf* the Medline-CD) for each one the study's specialisations

BioMedNet	Observed agreement (%)	+/+ agreement	-/- agreement	Kappa	P value
Internal medicine	63,1	30,3	32,9	0.3320	0.0000
Pharmacy	55,5	20,7	34,8	0.1682	0.0000
Gynaecology-obstetrics	62,7	27,2	35,5	0.2488	0.0000
Surgery	60,3	38,5	21,7	0.2109	0.0000
Paediatrics	78,1	20,1	58.0	0.4893	0.0000

DISCUSSION

Doctors frequently use books edited by experts concerning specific themes as their sources for consultation to keep abreast of current developments in their fields; they also have to read specialised journals. As well as being laborious, this practice does not always generate answers to those questions posed; it is often divorced from clinical scenarios or those not easily found in journals read by doctors, meaning that it cannot be readily applied to resolving questions concerning patients (14). Summaries have thus been produced in physical medium permitting manual searches of information available during a particular year concerning a specific theme (15). This tool has become popular with the availability of mass-media information, now available magnetically (one of these being Medline).

The Internet network has allowed numerous users to have instantaneous access to information (which can be immediately available), making several free-access servers available in Internet for searching medical databases (i.e. Medline). A lot of questionable quality information is dispersed through the network and periodical publications so that doctors are subjected to vast amounts of information which is of very little use, thus facilitating EBM's arrival onto the scene. This represents an attempt to ensure that all diagnostic, prognostic and therapeutical decisions should be based on solid numerical tests proceeding from the best clinical-epidemiological research (16). Using EBM is still not simple as training is required in such process, even more so when many obstacles are found in obtaining information, such as the type of source to be consulted, where such sources can be found, the order of consulting sources and problems related to handling computers (17).

Several experts have concluded (considering that regulating information available on the Internet is extremely difficult) that it is necessary to train or guide a consumer in finding the most reliable information available on the network (18). Other authors have focused more on the consumer than the source of information itself by carrying out studies regarding consumer trends concerning those sources of information which they consult, the way it is consulted and the tools used (19-21).

Medline-CD's sensitivity for detecting articles has already been evaluated when compared against a gold standard (i.e. a manual search). It was determined that the Medline-CD had 82 % sensitivity in 1988 when a manual search was made of all ophthalmology journals and 87 % when the gold standard was a manual search of indexed journals in the same specialisation (22). This revealed an important problem when defining the gold standard for this type of study and how a particular test's operational characteristics change when defining it is discretely varied. This study was designed on agreement, given that the Medline-CD could not be considered as being a gold standard in itself (based on the results of such evaluation).

Levels of agreement with the Medline-CD estimated in this work were low when employing search engines commonly used by doctors. This was worrying, considering that the scenario dealt with the Medline-CD having 82% sensitivity when compared to manual search and that error between available search engines and through Internet and the Medline-CD must be added to this error. Searches generally showed better behaviour for BioMedNet regarding specialisation, but at the expense of greater -/- agreement (Table 5).

+/+ rather than -/- agreement is preferable for a doctor. Not obtaining the information needed to answer his/her question (i.e. greater -/- agreement) due to poor selection of search engine can result in not finding articles relevant to a correct, systematic answer. It is thus assumed that finding greater +/+ than -/- agreement is preferable, even though general agreement may be high. If a reference for comparison (in this case Medline-CD) is responsible for low agreement (and not the search engines which are being evaluated), then having low agreement is attributed to just the search engine, when responsibility should really be shared.

Greater +/+ and -/- agreement can be used for evaluating information offered by search engines. The BioMedNet search engine had greater -/- agreement than the other search engines, even though it presented high agreement respecting the Medline-CD; thus, although its agreement might have been high, its performance would not benefit a doctor interested in +/+ agreement.

Given the information obtained through this study, agreement values were seen to be very low (54,1 % - 65,8 %) respecting those desired (greater than 85 %) (23). This is even more worrying when it is supposed that accessing the same database by different routes supposes that the same information will be found when using identical strategies. It thus becomes important to evaluate the quality of the software produced and/or maintained by webpages providing this service when determining whether the lack of results' reproducibility has been produced by their own inherent characteristics. It was observed that search engines have different behaviour (in terms of agreement) regarding each specialisation. The Medscape search engine (generally presenting 54,8 % agreement) presented the highest gynaecology-obstetrics value (75,9 %).

Servers' advanced options refined a search by offering limiters and suggested indications for obtaining better results. Such options are very useful if the theme to be sought is broadly known and the operator has gained experience from making previous searches, enabling more sensitive strategies to be employed for a particular theme (as done and designed in this project). Laymen generally do not use these tools correctly in daily life, carrying out searches limited by just temporal aspects.

Limitations

It is probable that excluding search strategies having more than 100 results might also have excluded results having better +/+ agreement which could have improved the results' final agreement. However, given the conditions of real life, a search strategy generating such a quantity of articles for a doctor to review becomes practically inapplicable. Additionally, selective exclusion of articles could have affected Medscape and BioMedNet search agreement, given that such searches were excluded for having more than 100 results in the same search, increasing the possibility that there could have been more agreement with the Medline-CD. However, these two search engines presented the best behaviour. These results cannot be generalised to other search engines, other operators, nor can they be compared with those resulting from a manual search.

Conclusions

Even though the availability of free access to Medline has favoured the boom in EBM, the results of this work must be considered within the context of choosing the access route for making the searches. Internet search engines revealed poor agreement; such agreement was differential according to specialisation, prompting poor reproducibility when using these search engines. It is thus possible that constructing in-house software for each search engine significantly affected results produced by whatever type of search process in the intent to find answers to questions related to the medical profession.

The results thus show that when making a general search, having no limitations (except temporal ones) and using correctly carried out search strategies, then the search engines having the best general agreement were BioMedNet and Gateway. BioMedNet had better -/- agreement than the other search engines and Gateway had better +/+ agreement. Given these results, new search engines' maintenance and validation processes must become standardised in the future to know *a priori* the expected level of search agreement, as the error expected for the Medline-CD (against a manual search) as well as the error committed by the search engine in selecting the information must be added.

The most concordant search engine will have to be used for now, whilst the process for creating new search engines having a less error is being standardised. It is consequently left in the doctors' hands to reflect on what they want from a search process (i.e. whether they wish to use a search engine having greater general agreement, greater +/+ agreement or greater -/- agreement) •

Acknowledgements. We would like to thank Hernando Gaitán, Rodrigo Pardo and Jason Garry for revising the manuscript and Jairo Echeverry for providing methodological support. We should also like to thank those people in PubMed (Eugenia Posey-Marcos), BioMedNet (Mark Merifield) and Gateway customer services (Margaret McGhee), for providing information about their respective search software.

REFERENCES

1. Wilson P. How to find the good and avoid the bad or ugly: a short guide to tools for rating quality of health information on the internet. BMJ. 2002;324:598–602

2. Bowden VM, Kromer ME, and Tobia RC. Assessment of physicians' information needs in five Texas counties. Bull Med Libr Assoc. 1994 Apr; 82(2):189–96.
3. Modlin M. Medical questions? Medline has answers. American Libraries. 1999; 29(10): 40-43.
4. Haynes RB, Wilczynski NL, McKibbon KA, Walker-Dilks CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc. 1994; 1: 447 – 458
5. Felix Free Medline Page. Updated 6th October, 1998. [Internet]. Available at http://www.akdeniz.edu.tr/tip/gcerrahi/drfelix. Visited in September, 2001 at http://www.beaker.iupui.edu/drfelix
6. Self PC, Sayed EN, Henry JK. Bridging the information gap" for Virginia public health nurses. Public Health Nurs. 1997 Jun;14(3):151-5.
7. National Library of Medicine, PubMed. [Internet] Available at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed. Visited in September, 2001
8. National Library of Medicine, Gateway. [Internet]. Available at http://gateway.nlm.nih.gov/gw/Cmd. Visited in October, 2001
9. WebMD Medscape Health Network, Medline. [Internet]. Available at http://www.medscape.com/px/urlinfo. Visited in October, 2001
10. ScienceDirect®, trademark of Elsevier B.V., Medline. [Internet]. Available at http://www.bmn.com/. Visited in September, 2001
11. Silver Platter International N.V. Adobe Systems Incorporated. 1991-1999.
12. Sackett D, Straus S, Richardson S, Rosenberg W, Haynes R. Evidence Based Medicine. How to Practice and Teach. 2nd. Edition. New York: Churchill Livingstone. 2000. p.p. 29-65.
13. Pérez A, Rodríguez N, Fabián J, Ramírez G. Software: Tamaño de la Muestra. Versión 1.1. Pontificia Universidad Javeriana. Facultad de Medicina. Unidad de Epidemiología Clínica y Estadística.
14. Sackett D, Straus S, Richardson S, Rosenberg W, Haynes R. Evidence Based Medicine. How to Practice and Teach. 2nd. Edition. New York: Churchill Livingstone; 2000. p. 1-5.
15. Cumulated Index Medicus. Available only in print. National Library of Medicine. [Internet]. Available at http://www.nlm.nih.gov/tsd/serials/. Visited in October, 2001
16. Guyatt G. Preface. In Guyatt G, Rennie D. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. Chicago: American Medical Association, 2002. 736 p.
17. Ely JW, Osheroff JA, Ebell M, Chambliss ML, Vinson D, Stevermer J, Pifer E. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ. 2002;324:710.
18. Purcell G, Wilson P, Delamothe T. The quality of health information on the Internet. BMJ. 2002; 324: 557-558.
19. Gagliardi A, Jadad AR. Examination of instruments used to rate quality of health information on the Internet: chronicle of a voyage with an unclear destination. BMJ. 2002; 324: 569-573.

20. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests and in-depth interviews. BMJ. 2002; 324: 573-577.
21. Wilson P, Risk A. How to find the good and avoid the bad or ugly: a short guide to tools for rating quality of health information on the Internet • Commentary: On the way to quality. BMJ. 2002; 324: 598-602.
22. Dikersin K, Scherer R, Lefebvre C. Systematic Reviews: identifying relevant studies for systematic reviews. BMJ. 1994;309: 1286-1291.
23. Streiner D, Norman G. Health Measurement Scales. Oxford University Press Inc: New York; 1998. p.121.