

Access to Information on Medical Research and Health Care Issues Relating to Diverse Population Groups: a Study of Embase, Medline and PsychINFO.

Efthimis N. Efthimiadis¹ and Marianne Afifi²

1 Department of Library & Information Science, Graduate School of Education & Information Studies, UCLA, CA 90095, USA

2 Center for Scholarly Technology, University Library, USC, CA 90089, USA

1. Introduction

The increase in the number and variety of ethnic populations in the United States and the concerns about their health have led to an expansion of demand for research on those populations in medicine and other fields such as public health and social welfare. Health policy makers, planners and service providers are seeking a better understanding of the characteristics of ethnic and racial groups in order to address needs. Health workers need to be sensitive to the ethnocultural barriers that confront different patients and to be aware of disease distributions in diverse populations. Studying a variety of ethnic and racial groups could establish the nature of differences in these disease distributions and their risk factors, and consequently lead to improved prevention and treatment modes. Research-granting bodies have already recognized the need for better understanding in these areas and many steps have been taken to further this goal (Department of Health and Human Services, 1991).

Parallel to this much needed research, we must examine the methods that facilitate access to the research literature. Research materials and findings are useless unless they can be accessed and properly disseminated. Current indexing procedures followed by the leading database producers do not adequately fulfill the needs of the research community concerning specific ethnic or racial groups. Similar inadequacies in indexing procedures are found in other Abstracting & Indexing Services. The main goal of the research presented in this paper is to investigate the methods of access (indexing and retrieval) to medical research on population groups in the three major health science databases: Medline, Embase, and PsycINFO.

2. Literature Review

The literature pertaining to access of information relating to diverse population groups focuses on three main areas that deal with (a) the confusion about the definitions of the terms race and ethnicity, (b) the correct and incorrect use of terminology, and (c) the lack of research data on racial and ethnic groups.

Many of the complications encountered in retrieving research data about ethnic populations lie in the definitions of the terms race and ethnicity. Sheldon and Parker (1992) discuss the lack of terminological and measurement consistency in the medical literature. They point out to the confusion between the concepts of race (biological) and ethnicity (cultural) and the lack of consistency in their operationalization". A possible reason for this confusion could be a reluctance to use the term race and substitution of the term ethnicity to appear politically correct. Thus the two terms are often used interchangeably and incorrectly.

In data collection and analysis, race and ethnic categories are often used in an all-inclusive manner, making no distinction between subgroups included in a given category. See for example the discussions of Yu and Liu (1992) on Asians and Pacific Islanders, Bhopal, Phillimore and Kholi (1991) on Asians as pertain to the Indian subcontinent, and Hayes-Bautista and Chapa (1987) and Treviño (1987) on Hispanics and Latinos, respectively.

The report *Healthy People 2000: National Health Promotion and Disease Prevention Objectives* (Department of Health and Human Services, 1991) emphasizes that data are especially limited for special populations such as racial

and ethnic minority groups. In addition, the report recognizes the need for more detailed information-gathering and dissemination on ethnic populations. The literature also indicates that there is great need to improve access to existing and future data on ethnic and racial populations from both a medical and a public health perspective. From a clinical practice perspective, access to research findings is essential to improving diagnosis and treatment modalities as discussed by Zhou et. al. (1989).

3. Research Objectives

The objectives of this research are (a) to study the indexing practices of the major secondary abstracting and indexing (A&I) services of the health sciences literature and (b) to determine the extent of access to ethnic and minority group data.

4. Methodology

4.1. Selection of A&I Services

The A&I services considered in this study are The National Library of Medicine's *Index Medicus* (Medline), the American Psychological Association's *Psychological Abstracts* (PsycINFO), and Elsevier's *Excerpta Medica* (Embase). These were selected because they are the major representative health science databases.

4.2. Examination of Thesauri

To achieve the above objectives and determine how ethnic and racial terms are represented in the three A&I services, the printed and online thesauri and indexing procedures for Medline, Embase, and PsycINFO were examined first. Indexing terms in use were identified and examined as to their adequate representation in the current literature. The medical literature contained in the three online databases was searched using both controlled vocabulary and natural language terms. A brief description of each thesaurus is presented here in order to establish the background on which the subsequent discussion is based.

4.2.1. MeSH

The Medical Subject Headings, (National Library of Medicine, 1993) allow for twelve ethnic groups. Scope notes provide instruction on how to use terms, which subheadings are allowed, and also indicate usage of other, older terms. Within the tree structures of MeSH, the various related terms were found under the following classifications:

1. Social Sciences - Anthropology - Anthropology, Cultural - Culture - Ethnic Groups (I1).

The term **Ethnic Groups** was further subdivided into the following twelve groups:

Aborigines, Asian Americans, Blacks, Eskimos, Gypsies, Hispanic Americans, Mexican Americans, Indians Central American, Indians North American, Indians South American, Jews, Whites.

2. Social Sciences - Anthropology - Anthropology, Physical - Racial Stocks (I1).

Under this classification the term **Racial Stocks** can be expanded to:

Australoid Race, Caucasoid Race, Mongoloid Race, Negroid Race

The same ethnic groups as above can be found in:

3. Social Sciences - Sociology - Culture - Ethnic Groups (I1); 4. Named Groups (Non-MeSH) - Ethnic Groups (M1); 5. Population Characteristics (Non-MeSH)- Demography - Ethnic Groups (N1); 6. Geographic Names (Z1) (for specific countries).

To describe other ethnic groups than the ones listed above, MeSH suggests to combine the **Geographic Names** tree structure together with the subheading **Ethnology**. For example, an article that describes particular disease patterns in Japanese populations would be indexed as **Japan-Ethnology**.

Thus the MeSH indexing provides only for a limited number of named ethnic groups. Although much of the research has been on some of these groups, others are neglected. The **Geographic Names** and their subheadings are inadequate, especially considering named ethnic groups within other countries and major ethnic groups in the United

States not named in the above categories. For example, the Hmong people, a minority tribe living on the border between Thailand and Laos, many of whom were resettled in the United States after the Vietnam War, are represented in the controlled vocabulary by the heading **Laos - Ethnology**.

4.2.2. P.I.T. Thesaurus

The Thesaurus of Psychological Index Terms (P.I.T.) provides coverage of the major ethnic groups and adequately relates terms not used to terms that are used. For example, if users look up the terms **Negroes** or **African American**, they are instructed to use the term **Blacks** instead. The term **Ethnic Groups** lists seven narrower terms: American Indian, Anglos, Arabs, Asians, Eskimos, Gypsies, Hispanics (Narrower Term - Mexican Americans).

There are ten related terms under the heading **Ethnic Groups**, among others the terms **Blacks** and **Whites**. Other than for the term **Mexican Americans**, the thesaurus does not provide narrower terms for the other ethnic groups. Scope notes are provided for terms such as **Asians**, which only includes people of Asian origin not residing in their original country of origin. Others are indexed under the country that they reside in. This distinction, although made clear in the scope note, may not be one that an average user would make. The term **Race (Anthropological)** also exists in the thesaurus with eight narrower terms, among them **Asians**, **Blacks**, and **Whites**. No scope notes are provided for the term.

The thesaurus features nine alphabetical clusters and subclusters which are intended to facilitate searching for novice users. According to the introductory material, these clusters are assembled on the basis of similarity and likeness. The only cluster of interest for this discussion is the **Geographic Cluster** which contains hierarchically arranged names of countries. Ethnic groups are not part of any of the clusters.

4.2.3. EMTREE

The EMTREE Thesaurus is based on tree structures called facets. It consists of 15 Facets (A-Q) each of which has subfacets beginning with the same letter and numbered sequentially, for example K1, K2 etc. The facets of interest for our purposes are **Facet K: Geographic Names** and **Facet M: Named groups of persons**. Facet K is further subdivided into four regions, within those into subregions, and then into individual countries. Facet M is divided into three subgroups: **M1: Ethnic and Racial Groups**; **M2: Race**; **M3: Religious Groups**.

M1 is further subdivided into five named groups. These are: **Aborigine**; **Ethnic Group**; **American Indian**; **Indian**; **Chinese**. **M2**, the Race classification includes the terms **Caucasian**, **Mongoloid Race**, and **Negro**, whereas the religious group classification **M3** contains the terms **Christian**, **Jew**, and **Moslem**. Each major facet is briefly described but no scope notes are provided. Thus, the thesaurus allows for limited indexing of terms for ethnic groups.

4.3. Searching the databases

After examining both the printed and online versions of Medline, PsycINFO, and Embase a list of 145 terms relating to ethnic and racial groups was compiled. To establish a usage history over time, each of the three databases was searched on Dialog in five year increments starting at their inception and ending with 1993 (see Table 1). The results allowed the comparison of the usage of each term over the specified periods for each database and between databases. Each term was searched 25 times for each of the time periods. Terms were searched without field limits, i.e., in Dialog's Basic Index (BI), and with limits, restricting the search to the descriptor (DE), title (TI), abstract (AB) and identifier (ID) fields respectively. The DE and ID fields representing the controlled vocabulary of the systems, whereas the title and abstract fields represent the vocabulary of the users (authors and searchers).

4.4. Selection of terms

From the search results 33 terms which were controlled vocabulary terms in use in at least one of the three databases, were selected for the detailed analysis that is reported in this research. For each term, a summary measure was computed based on the average occurrence of the term over the search period for each of the four fields, i.e., the Basic Index total (BI), the descriptor field (DE), the title field (TI), and the abstract field (AB). The same was obtained for DE, TI, and AB as a percentage of the total occurrences. The ID field was not included for the detailed analysis because Medline does not use an ID field, and Embase uses it sparingly. To determine trends, the slopes of the total occurrences, DE, TI, AB, DE%, TI%, and AB% were obtained as a function of time using the least

squares method. In the latter analysis, the consecutive five-year periods were coded as 5 for the latest period and 4, 3, 2, and 1 corresponding to the earlier periods. Since there were only four time periods for Embase, the earliest period was coded as 2, as seen in Table 1.

Table 1. Study periods for each database.

Medline	No.	PsycINFO	No.	Embase	No.
1988-1993	5	1988-1993	5	1988-1993	5
1983-1987	4	1983-1987	4	1983-1987	4
1978-1982	3	1978-1982	3	1978-1982	3
1973-1977	2	1973-1977	2	1974-1977	2
1966-1972	1	1967-1972	1		

4.5. Grouping of terms

The terms were further analyzed and were clustered into 9 groups. The grouping of the terms was intended to reflect an end user's perspective and for the most part it brings together terms that are associated with or about a population group. For example, the terms **Mongoloid Race**, **Asian/s**, **Chinese** and **Asian American/s** can be considered related terms by a searcher. Of the 9 groups, seven were identified containing interrelated terms. Group 8 included three terms that were difficult to fit in any other group, while Group 9 consists of a collection of general terms such as **Ethnicity**. Since such terms are not likely to be used by a searcher to identify a specific population group, Group 9 was omitted from further detailed analysis. The groups are discussed in the results section.

5. Results and Discussion

Due to space limitations the results are discussed but cannot be presented here in terms of tables and figures.

The three thesauri revealed shallow pre-coordinated hierarchical structures, rather difficult to use terms and post-coordination, and a blurring between cultural, genetic, and racial facets. A big area of concern is the use warrant, for example for the terms **Blacks**, **Hispanics**, and **Indians**. In addition, terminology which is scientifically outdated is used -- for example, the terms **Caucasoid**, **Mongoloid**, **Australoid**, and **Negroid** under **Racial Stocks** in the MeSH terms. Furthermore, post-coordination is difficult because of the system-oriented terminology which is mostly intended for information professionals.

The terminology unintentionally restricts access to the end users who lack the needed knowledge to use the thesaurus effectively for information retrieval. Returning to the example where an end user is seeking information on the **Hmong** people in Medline, for an inexperienced searcher who wishes to perform a controlled vocabulary search, nothing about the **Hmong** would appear in the descriptor field, i.e., the term **Hmong** retrieves nothing, because indexing is done under **Laos-Ethnology**. The end users would thus be misled and dissatisfied. Misled because they may think that there are no records on the subject; dissatisfied because they may know that there are articles on the subject but they cannot retrieve them. Therefore, except for trained information professionals, end users will have difficulties retrieving what they need.

5.1. Results by Group

Trends in total usage of the selected terms over time were generally positive. Terms showing a negative slope tended to be those denoting race and not ethnicity. The terms denoting ethnicity and showing a negative slope were likely to be outdated such as, the term **Negro**. These results indicate that the indexing policies have shifted towards using controlled vocabulary terms denoting ethnicity rather than race. The results for Groups 1 through 7 are presented here. Instead of looking at each term separately, the results are reviewed according to the groups of related terms

as defined in the Methodology.

5.1.1. Group 1 [Asian? ? (P); Asian()American? ? (M); Chinese (E); Mongoloid()Race (M)]

The term **Asian/s** shows increases in all three databases and in all measures (BI, DE, TI and AB). While it was not a descriptor for all periods in all three databases, increases are evident in Medline (M) and PsycINFO (P) for those periods where it was. Embase (E), being based in Europe and thus defining **Asian/s** differently than American database producers, only used the term as a descriptor in period 5. The term **Asian American/s** also increased in all measures, although the total number of records was relatively small. The percentage increase from period 4 to period 5 was high, especially in Medline and PsycINFO. The term was in use as a descriptor in Medline for periods 2-5, it was eliminated in PsycINFO in periods 4 and 5, and it was not a descriptor in Embase until period 5. The term **Chinese** showed a strong increase in all indexes searched over the span of the five periods in all databases, but especially so in Medline. However, not all the records appearing in the descriptor field in Medline are actually records describing Chinese people. There, the occurrence of the term denotes adjectives combined with nouns, such as Chinese medicine, while a typical user may interpret the term as a noun referring to people from China. The idea behind such distinction between adjectives and nouns is that documents referring to people from China will be classified as **China-Ethnology**. A similar situation occurs in Embase although there **Chinese** is a thesaurus term, while it is not in Medline. In PsycINFO, the term **Chinese** is only evident in the TI and AB fields. The term **Mongoloid Race** shows an overall decrease in the BI in Medline and it is almost exclusively used in the descriptor. The number of records, relative to other terms in the group, is small. The term was a descriptor in Embase for periods 4 and 5 but the number of records is negligible. The term was not found in PsycINFO.

5.1.2. Group 2 [Blacks (M,P); Negro or Negroes (E); Negroid()Race (M)]

The use of the term **Blacks** exhibited an increase over all five periods in all measures except for the descriptors in Medline and Embase. In Embase, the term was not a descriptor while in Medline there are increases in the descriptor from period 3 to period 5. PsycINFO shows a sudden increase in the number of records for periods 4 and 5 in the DE field. The terms **Negro or Negroes** were used as descriptors only in Embase and the data show a tenfold increase in the use of the term in period 5. In PsycINFO, the number of records increased from periods 1-2, then decreased by half in period 3 and is negligible for periods 4 and 5. In the TI and AB fields the use of the terms decreased significantly from period 1 to period 5. The sudden decrease in the terms in the BI may be related to the sudden increase in the term **Blacks** in the DE field. The term **Negroid Race** is in use only in Medline as a descriptor and shows a decrease from period 1 to period 3 and then a sharp increase in periods 4 and 5. There are negligible data in Embase and none in PsycINFO. Overall, the number of records for the term **Blacks** is much larger than for the others combined.

5.1.3. Group 3 [Hispanic? (P); Hispanic()American (M); Mexican American (M,P)]

The term **Hispanic?** shows an overall increase in all databases in all measured fields, except for the descriptor in Embase. The term is a descriptor in PsycINFO, but not in the other databases, even though there are occurrences of the term in the DE field in both Medline and Embase. The term **Hispanic American??** is a descriptor in Medline but not in the other databases. It shows an increase over the five periods in all databases especially from period 4 to 5. It is noteworthy that the number of records in the DE field is fairly high compared to the TI and AB fields. The term **Mexican American?** was used as a descriptor in Medline only in period 5, is a descriptor in PsycINFO in all periods, and is not a descriptor in Embase. The data in Medline in the TI and AB fields show an increase over the five periods which is strongest in the last two periods. The data in PsycINFO exhibit their strongest increase in period 3 for all fields except for the AB field where the strongest increase is in period 4. The TI and AB fields show an overall increase whereas the DE field shows a slight decline in period five.

5.1.4. Group 4 [American()Indian (E,P); Indian? (E); Indians()Central()American (M); Indians(North()American (M); Indians()South()American (M)]

The terms **Indians-North American**, **Indians-Central American**, and **Indians-South American** are only used as descriptor terms in Medline. There are no data in the other databases and none in the TI and AB fields. It is evident that these terms are not realistic search terms because they appear only as descriptors. The term **Indian?** is only a descriptor term in PsycINFO (**Indian**), although data appear in the DE fields for both Medline and Embase because

of the use of the unrestricted truncation “?” in the search. In Medline, the data in the descriptor field include the other 3 search terms in this group but also many false drops such as Indian Ocean. In Embase, the descriptor term **Indian** refers to East Indians not Native Americans.

5.1.5. Group 5 [Anglo? (P); Caucasian? (E,P); Caucasoid()Race (M); Whites (M,P)]

The term **Anglo?** is a descriptor only in PsycINFO for periods 4 and 5. In the other fields the data show an upward trend. The number of records in each database and each field is less than 500. The term **Caucasian?** is a descriptor in Embase and was used as a descriptor in PsycINFO for periods 1 through 3. In Medline, the AB field shows a strong increase over all periods, while in the TI field, there is a moderate increase. In Embase, the numbers for the BI, TI, and AB fields show a strong increase from period 4 to 5. The term **Caucasoid Race** is used as a descriptor only in Medline. The number of records decreased from period 1 to 3 and then increased again in periods 4 and 5. The term **Whites** is used as a descriptor in Medline and PsycINFO. In Medline and Embase, all fields for which there are data exhibit a rise in the number of records, especially from period 4 to 5, although the number of records for the TI field is relatively small compared to the AB field. In PsycINFO, the number of records is largest in the DE field, where the strongest increase was from period 3 to 4. The data in the TI and AB fields are variable.

5.1.6. Group 6 [Eskimo? (M,P); Inuit (M)]

The term **Eskimo?** is a descriptor in all three databases. In Medline, the number of records decreased from period 1 to 3 and increased again in periods 4 and 5. The number of records in the TI field decreased steadily from period 1 to period 5, while in the AB field there is an increase over these periods. In Embase, there are increases in the DE field overall, in the TI and AB fields there are overall slight decreases in the number of records. A similar pattern has been observed in the PsycINFO data. The total number of records for any one period in all databases is less than 400. The term **Inuit** is not a descriptor for any of the databases. The number of records increases steadily in Medline and Embase in the other fields, while in PsycINFO the data are variable.

5.1.7. Group 7 [Aborigine? (E,M); Australoid()Race (M)]

The term **Aborigine** is a descriptor term in Medline (period 5 only) and in Embase for all periods. Although the term is colloquially used for the aborigines of Australia, the online Medline thesaurus defines aborigines as “native inhabitants or indigenous individuals of a country”. Approximately three quarters of the retrieved records in Medline also contain the term **Australoid Race** in the descriptor or contain the geographical term Australia thus referring to Australian aborigines. The number of records in Medline and Embase shows strong increases from period 4 to period 5 in all fields and are small and steady in PsycINFO. **Australoid Race** is a descriptor in Medline and the only retrievals for the term are in the DE field. The number of records is fairly steady over all periods and decreased just slightly overall.

6. Concluding Remarks

Overall, there is a general and gradual increase beginning in 1966 in the terms relating to race and ethnicity in all three databases over the periods examined. These trends show an increased interest in the health of specific population groups. Regardless of the reasons for the increase, it is clear that the researchers’ needs are not well served by the databases. Some terms are found in some databases, but not in others, some are introduced into a database while being dropped from another. This confusing state of affairs may be explained, but only in part, by the fact that the “political correctness” of the terms changes over time. Indeed, when a database includes several more-or-less synonymous terms, there is a gradual shift in occurrence from some terms to others. Notwithstanding this shift in usage, the databases are clearly not in agreement with regard to the definition or use of most of the terms examined.

Based on the findings, the following areas have been identified that need to be addressed by A&I services:

- a) The indexing process needs rethinking to accommodate medical and public health research concerns.
- b) All users of these databases, be they information professionals, health professionals, students, or lay persons, need to be alerted to the difficulties that may be encountered in searching. Searchers may not be able to access the literature if they are not familiar with the indexing policies of population groups.
- c) It is difficult to ask that medical researchers be familiar with the indexing policies. Health sciences information

professionals may be required to have exceptional familiarity with the databases and the indexing, but it certainly cannot be expected of any other searchers. Therefore, the database providers may want to look at their procedures and adapt them in such a way that an inexperienced searcher can find the appropriate literature on ethnic population groups without extended knowledge of the indexing.

d) It is well known that indexer inconsistency contributes to faulty indexing, that is, two indexers may use a different choice of terms to index the same document. The cause of these inconsistencies is likely human error, but ultimately the indexing policy may contribute to the problem.

Therefore, there needs to be increased sensitivity on the part of the database producers to pay attention to the current trends in the research of population groups. In addition to an expansion of the number of thesaurus terms for population groups, additional indexing that would benefit medical researchers needs to be considered. Some attributes for further indexing may include the type of study, sample sizes and age groups.

7. References

1. Bhopal, R. S., Phillimore, P., & Kholi, H. S. (1991). Inappropriate use of the term 'Asian': an obstacle to ethnicity and health research. Journal of Public Health Medicine, 13(4), 244-246.
2. Department of Health and Human Services. (1991). Healthy People 2000: National Health Promotion and Disease Prevention Objectives, (DHHS Publication No. (PHS) 91-50212). Washington, DC: U.S. Government Printing Office.
3. Excerpta Medica. EMTREE Thesaurus. (1992). New York, NY: Elsevier Science Publishers.
4. Hayes-Bautista D. E., & Chapa J. (1987). Latino Terminology: Conceptual Bases for Standardized Terminology. American Journal of Public Health, 77(1), 61-69.
5. National Library of Medicine. (1993). Medical Subject Headings: Annotated Alphabetic List; and Tree structures.
6. National Library of Medicine. (1987). Indexing Manual. Section 30, Category I-2.
7. Sheldon, T. A., & Parker H. (1992). Race and ethnicity in health research. Journal of Public Health Medicine, 14(2), 104-110.
8. Trevino, F. M. (1987). Standardized Terminology for Hispanic Populations. American Journal of Public Health, 77(1), 69-72.
9. Walker, A., Jr. (Ed.). (1994). Thesaurus of psychological index terms. Washington, D.C. : American Psychological Association.
10. Yu, E. S. H., & Liu, W. T. (1992). US National Health Data on Asian Americans and Pacific Islanders: A Research Agenda for the 1990s. American Journal of Public Health, 82(12), 1645-1652.
11. Zhou, H. H., Koshakji, R. P., Silberstein, D. J., Wilkinson, G. R., & Wood, A. J. J. (1989). Racial Differences in Drug Response. The New England Journal of Medicine, 320(9), 565-570.