Mendelian Inheritance in Man: Diagnoses in the UMLS

by Karen M. O'Keefe, MaryEllen Sievert, Ph.D., and Joyce A. Mitchell, Ph.D.
Medical Informatics Group, University of Missouri-Columbia
School of Library and Informational Science, University of Missouri-Columbia

ABSTRACT

Because they deal with many distinct but rare inheritance diseases, geneticists have difficulty translating from their codes to other biomedical coding schemes. The objective of this research was to investigate the potential uses and difficulties of using the UMLS Metathesaurus for genetic diagnoses and to make recommendations to UMLS developers for improvements in UMLS for common genetic disorders. The 110 most common Mendelian Inheritance in Man disorders from the Missouri Genetic Disease Program over the period of one year were translated into MeSH, ICD and SNOMED. The more common diseases are more likely to be mapped than the rarer ones. Diseases with a proven genetic inheritance pattern are more likely to be mapped than those with speculated inheritance patterns. Approximately one third of all diagnoses were not mapped across all three coding schemes in Meta-1.2. The ICD coding scheme was found to be too broad to be meaningful for genetic diagnosis or epidemiological purposes. MeSH and SNOMED need to be made more specific and complete, and all of the new version of SNOMED needs to be included in the Metathesaurus.

INTRODUCTION

The researchers' intent was to investigate the potential helpfulness and difficulties of using the UMLS Metathesaurus for genetic diagnoses and to make recommendations to its developers in refining UMLS for common genetic disorders. Translating between the many diverse codes is a problem for geneticists. Geneticists must translate their diagnoses to MeSH to search the medical literature, to SNOMED to consult with other medical specialists, to ICD and to CPT to do patient billing via insurance companies, and to ICD to do epidemiological reporting.

Another objective of this research project was to evaluate the quality of genetics mapping data using various coding schemes and to make recommendations for genetics reporting. Geneticists tend to be especially frustrated with ICD codes, since they are very broad, often including many genetic diseases, as well as those of other etiologies, into one broad heading, such as "Hearing Loss". Because epidemiological reporting is currently done with ICD codes, geneticists have a very vague idea about the incidence of specific genetic diseases on a nation-wide level.

This project focused on one of the many types of genetic diseases, Mendelian inheritance, as described and coded in Mendelian Inheritance in Man (MIM)[1]. The specific project research addressed which of the approximately one hundred most common diagnoses could be found using Meta-1.2, the most current version of the UMLS system. If a diagnosis was found in Meta-1.2, research addressed whether it existed as a preferred term in the Meta-1.2 coded portion of ICD, of SNOMED and of MeSH. The diagnosis was translated into the preferred ICD, SNOMED, and MeSH terms and codes, if available.

REVIEW OF LITERATURE

A major function of the UMLS Metathesaurus is to translate between different biomedical coding schemes. MIM and OMIM, the online version of MIM, cover all known dominant, recessive and X-linked traits. Entries with an asterisk represent proven inheritance patterns by a specific gene at a unique locus while the inheritance pattern of unstarred entries is not proven to be associated with a specific gene [1].

Many diverse coding schemes including MeSH, ICD, and SNOMED, are interrelated via UMLS. The objective of most users is, given an entry term or even a preferred term in one coding scheme, to find a preferred term for one or more other coding schemes. Other articles have discussed the use of the Metathesaurus [2,3] but these have not looked at genetic diseases. Tuttle writes of the problem of missing terms in Meta-1.1. He predicts that "even a perfunctory analysis of 'missing' terms at most sites will show a strong preponderance of 'precoordinated' terms each of whose constituent parts are already in Meta-1.1" [4]. The evaluation criteria of this research project has been designed to examine this hypothesis about precoordinated terms with regard to MIM diagnoses.

METHODOLOGY

The research project matched MIM diagnoses

0195-4210/92/$5.00

and codes with the terminology and codes in other languages using the UMLS's Metathesaurus, Meta-1.2, released in November 1992. The hypercard browser version of Meta-1.2, mounted on CD-ROM, was used on a Macintosh IIci computer.

The sample selected consisted of those diagnoses coded with MIM numbers reported to the Missouri Genetic Disease Program having a prevalence of 2 cases or more for the fiscal year 1991-1992. These diagnoses were accumulated from the reports of 4 genetic diagnostic centers treating cases in different parts of Missouri, on the GOAS database, an online database for representation of genetic diseases in Missouri [5]. The MIM diagnoses having a prevalence of 2 or greater, 110 diagnoses,were selected from this larger sample of 333 diagnoses representing 951 individual patients.

The procedure for comparing the schemes was as follows: First the preferred term for each disease was chosen as an entry term to the Metathesaurus. If that term was not in the Metathesaurus, every other synonym, as found in the MIM catalog [1], The Birth Defects Encyclopedia [6], or Jablonski's Dictionary of Syndromes and Eponymic Diseases [7], was tried as an entry term. Once an entry term was found in Meta-1.2, that term's concept card was opened. The coding numbers in the Source field were noted for each coding scheme. In the terms card for that concept, we looked for lexical variants, synonyms or reviewed related terms as a preferred term with that coding number.

After all preferred terms for each coding scheme were found, the quality of the match between each coding scheme and MIM was considered. The possible categories of a match were as follows:

(1) An Exact Match

Exact matches were scored when the preferred term in a particular coding scheme was a synonym for the specific preferred MIM term, as found in MIM, or the other reference sources, or whether the preferred term was a lexical variant of the MIM preferred term, meaning it had a different word ending or word order.

(2) Coding Scheme Broader than MIM Diagnosis

Broader terms consisted of terms for diseases with synonyms in the above mentioned reference sources but without the particular disease subtype, such as Mucopolysaccharidosis instead of Mucopolysaccharidosis IIIA. Broader terms also consisted of terms for concepts for more than one disease, or terms where a modifier such as congenital or familial was left off of the disease term. A broader term was also found whenever a symbol for a broader concept [B], existed in Meta-1.2 next to a term.

(3) Coding Scheme Narrower than a MIM Diagnosis

Narrower terms consisted of synonyms with additional modifiers or additional typing of the disease, or where the Meta-1.2 narrower symbol [N] appeared next to the synonym or reviewed related term.

(4) Coding Scheme Contains Two or More Terms Mapped to a MIM Diagnosis

Terms were considered to be mapped to two or more terms in a particular coding scheme if the source field contained two preferred codes.

(5) Coding Scheme Contains All Component Parts of a MIM Diagnosis

Complete precoordinated matches were found when all component terms of a disease were present as preferred terms in that coding scheme, but the entire disease, or its synonyms, were not present as a preferred term.

(6) Coding Scheme Contains Some, But Not All, Component Parts of a MIM Diagnosis

Incomplete precoordinated matches were found when some, but not all, component parts of a disease were present.

(7) MIM Term Not Mapped in Meta 1.2

Terms were considered not to be mapped when a coding scheme was not in the source field for an entry term for the specific MIM diagnosis, and when none of the synonyms, lexical variants or reviewed related terms were entry terms into that coding scheme.

Two examples illustrate the differences in the three schemes. For the MIM term "Pierre Robin Syndrome," MeSH uses the same term but in ICD, the term falls under "congenital anomalies of skull and face" (a broader term) and in SNOMED it is "Micrognathia-glossoptosis syndrome," an exact match with one of the synonyms. The second example, "Rett Syndrome," is again an exact match with the MeSH term; in ICD, it is under "other specified cerebral degenerations in childhood," a broader term; in SNOMED, there is no match.

Once it was known how many of the 110 diagnoses fit into each of the predefined seven categories, general patterns in the distribution of diagnoses were examined. The percentages of MIM diagnoses that fit into each of the categories mentioned in methods were calculated and Chi-squared tests were done where appropriate. Percentages of matches for each category were analyzed for trends that reflect usefulness of the matches to the medical geneticist.

RESULTS

Table I shows the number of diagnoses which fit into one of the predefined categories. (*Some of the SNOMED diagnoses in each category were mapped to the Morphology or Function Axes. Thus while their terminology was categorized by the researchers as an exact, a broader or a narrower match, they are not truly an equivalent match because they represent a match from a disease category in MIM to a non-disease category in SNOMED.)

Table I: Number of MIM terms Translated to Other Coding Schemes Geneticists Use

	MIM-MeSH	MIM-SNM *	,MIM-ICD
Exact match	53	39	24
Broader	22	24	37
Narrower	0	1	0
Not mapped	29	41	38
Incomplete Component Match	6	5	6
2 or more Matches	0	0	5
Total	110	110	110

Over a third of the MIM diagnoses in our experimental sample were not mapped either to SNOMED or ICD, or were mapped to neither. Almost a third of the MIM diagnoses were not mapped to MeSH.

Tables II, III, and IV compare exact matches and not mapped disorders, with and without asterisks, from MIM to MeSH, ICD and SNOMED respectively. In all cases incomplete component matches are considered as not mapped since they are not effectively mapped. A Chi-square test was calculated for each set of data.

Table II: MIM-MeSH Mappings: Comparison of Exact Matches and Not Mapped, With and Without Asterisks

	With*	Not *	Total
Exact Match	42	11	53
Not Mapped	22	13	35
Total	64	24	88

Table II compares exact matches and not mapped disorders, with and without asterisks, from MIM to MeSH. A Chi-squared test was done on this table with Chi-squared = 2.8542 which is statistically significant at the p=.10 level but not at the .05 level. This means that those disorders with proven inheritance patterns are somewhat more likely to be mapped in MeSH. It should be noted that all of MeSH is contained in Meta-1.2.

Table III: MIM to ICD Mappings: Comparison of Exact Matches and Not Mapped, With and Without Asterisks

	With*	Not *	Total
Exact Match	20	4	24
Not Mapped	28	16	44
Total	48	20	68

Table III shows the relationship between exact matches and not mapped disorders, with and without asterisks. A Chi-squared test was done on this table showing Chi-squared to be = 2.9026 which is statistically significant at the p=.10 level but not at the p=.05 level. This means that diagnoses with proven inheritance patterns are somewhat more likely to be mapped in ICD than those without asterisks. This finding must be considered in light of the fact that not all of ICD is contained in Meta-1.2.

Table IV compares exact matches and not mapped disorders, with and without asterisks, from MIM to SNOMED. A Chi-squared test was done on this table with Chi-squared = 4.9764 making the relationship statistically significant at the p=.05 level. This means that diagnoses with proven inheritance patterns are more likely to be mapped in SNOMED.

Table IV: MIM to SNM Mappings: Comparison of Exact Matches and Not Mapped With/Without Asterisks

	With *	Not *	Total
Exact Match	33	6	39
Not Mapped	29	17	46
Total	63	23	85

The relationship between those diagnoses with (32) and those without (4) an asterisk mapped to the Disease Axis in SNOMED were compared to those diagnoses with (24) and without (5) an asterisk

mapped to other axes, which were also compared with those diagnoses with (29) and without (17) an asterisk that were not mapped. A Chi-squared test showed Chi-squared = 8.3574. Thus the differences are statistically significant for p=.05. Thus disorders with proven inheritance patterns are more likely to be mapped to the Disease Axis in SNOMED and less likely not to be mapped.

Table V shows the relationship between frequency and inclusion in Meta-1.2. Those considered mapped to Meta-1.2 were mapped to one or more of the three studied coding languages. Those diagnoses considered not mapped to Meta-1.2 were not mapped to any of the three studied coding languages. Incomplete component matches were considered to not be effectively mapped to any of the three coding languages. A Chi-squared test for this table showed Chi squared to be 6.5506. The null hypothesis can be rejected for p=.050 and the differences are statistically significant. Thus more common disorders are more likely to be mapped than less common disorders.

Table V: Relationship between Frequency of Occurrence and Inclusion in Meta-1.2

Frequency	# Diagnoses Mapped	# Diagnoses Not Mapped	Total
> = 4 cases/ diagnosis	34	7	41
< 4 cases/ diagnosis	42	27	69
Total	76	34	110

DISCUSSION

While it can be seen from Table V that the more common diseases are more likely to be mapped to one of the coding languages than the rarer diseases, the geneticist deals in large part with the rarer diseases and needs to have those disorders available in other coding languages. The diseases showing a frequency of 2 or more made up only one third of the presenting diagnoses. In the same way, it can be seen from Tables II, III, and IV that disorders in MIM with a proven genetic inheritance pattern at a unique genetic locus were mapped to exact or broader matches more often than those without. While the majority of the cases seen by the geneticist are those that have proven inheritance, the disorders whose inheritance pattern are not yet proven are also an important part of the geneticists work. Both types of disorders asterisks showed large numbers that were not mapped in all three coding languages.

Almost a third (31.8%) of MIM diagnoses were not mapped to MeSH, yet all of MeSH is included in the Metathesaurus. The majority of SNOMED is included in Meta-1.2, yet 40% of the MIM diagnoses studied were not included in SNOMED. Another 44.6% of those mapped to SNOMED were not mapped to the Disease Axis. The majority of exact matches were mapped to the Disease axis while the majority of broader matches were mapped to the Morphology or Function Axes. Thus the failure to map certain diagnoses may not be the fault of the Metathesaurus, but of the MeSH and SNOMED coding languages themselves. Both languages should be made more complete, and the SNOMED coding scheme should include codes on the disease axis for all disease diagnoses.

From the examples, it can be seen that many of the ICD codes are broader in nature than the MIM codes. Often terms combine large classes of genetic diseases or combine acquired and genetic diseases. It can be concluded that ICD codes are, taken as a whole, less useful than other codes such as MIM, SNOMED and MeSH for genetics reporting.

Table IV demonstrates that there are several broader matches to MeSH and SNOMED. While this is still a significant problem for the geneticist, it can be seen from the examples that the matches are not so much broader, nor are there so many broader matches as to render the coding schemes useless for practical purposes.

What exactly is to be expected of the various coding schemes? MeSH and SNOMED, as seen by the large numbers of exact matches to MIM in each coding scheme have an acceptable degree of granularity to their coding schemes; what is most distressing is the many disorders that are not mapped at all. Including SNOMED in its entirety in the next edition of the Metathesaurus should partially resolve this problem. Beyond that, it would be helpful for future editions of MeSH to devote greater attention to expanding the scope of genetic diseases listed. This would not mean revising the coding scheme, only adding more terms to increase its scope. On the other hand, the number and kind of broader matches from MIM to ICD indicate that ICD does not have sufficient granularity to be useful to the geneticist in epidemiological reporting. Great revision would be required to make the granularity of ICD suitable for genetic diseases, As a result, geneticists are not well served by this method of epidemiological reporting. Including all of the new edition of SNOMED, expanding the scope of MeSH in the area of clinical

genetics, and adding all of the MIM disorders would greatly aid the field of clinical genetics.

We tested Tuttle's hypothesis that all precoordinated terms would be included in the Metathesaurus in their component forms [4]. While this may be a helpful explanation of mappings in other medical fields, we did not find it to be a helpful way of attempted matching for the geneticist. Many genetic syndromes contain as many as ten or more identifying characteristics and the major genetics resources do not agree on which set of identifying characteristics specify the disorder. Furthermore, a group of identifying characteristics, even when together they make up the preferred name of the disorder, does not indicate a well-recognized inheritance pattern. We did attempt to break down a few of the less elusive, otherwise not mapped precoordinated terms into component parts and found no complete matchings, only incomplete component matchings. In general, this hypothesis was not found to be useful to the geneticist. By making both the Metathesaurus and its composite coding schemes more specific and yet more comprehensive the geneticist and the medical community as a whole will be better served.

Acknowledgment: The research was supported in part by NIH grant LM07089 from the National Library of Medicine and by contract MDH-A0C3-00048 from the Missouri Dept. of Health.

References

1. McKusick, Victor. Mendelian Inheritance in Man: Catalogs of autosomal dominant, autosomal recessive and x-linked phenotype. 9th ed. Baltimore and London: John Hopkins University Press, 1990.

2. Cimino J.J. and Sideli R.V., "Using the UMLS to Bring the Library to the Bedside", Med Decis Making , 1991; 11:S116-9.

3. Chute C.G., Yang Y., Tuttle M.S., Sherertz D.D., Olson N.E., and Erlbaum M.S., MD, "A Preliminary Evaluation of the UMLS Metathesaurus for Patient Record Classification", SCAMC 1990, P.161-165.

4. Tuttle, M.S., Sherertz D.D., Erlbaum M.S., MD, Sperzel W.D., MD, Fuller L.F., Ph.D., and Olson N.E., Nelson S.J., MD, Cimino, J.J, MD, Chute, C.G., MD, DrPH, "Adding Your Terms and Relationships to the UMLS Metathesaurus" SCAMC 1992, P.219-23.

5.. Mitchell JA, Ph.D., Cutts JH, MS, and Hess, M, MS, "Use of a Microcomputer Database System in a Statewide Effort for Data Collection in Medical Genetics", SCAMC 1992:771-5.

6. Buyse, Mary Louise, MD, editor in chief, Birth Defects Encyclopedia: The comprehensive, systematic, illustrated reference source for the diagnosis, delineation, etiology, biodynamics, occurrence, prevention and treatment of human anomalies of clinical reference, Dover, Mass.: Center for Birth Defects Information Services Inc., 1987.

7. Jablonski, Stanley, Jablonski's Dictionary of Syndromes & Eponymic Diseases, 2nd ed., Malabar, Fla.: Krieger Publishing Co. 1991.