DIAGNOSTIC CODES IN MSP CLAIM DATA

Summary Report

Weimin Hu

Program Monitoring and Information Management Branch
Resource Management Division
Medical Services Plan

July 1996

Executive Summary

This report summarizes the results from an extensive study of the quality of diagnostic information recorded in the physicians' claims database (PMEMASTER) of the British Columbia Medical Services Plan (MSP), Ministry of Health and Ministry Responsible for Seniors. The objectives of this study are firstly to identify the various types of the diagnostic codes in the PMEMASTER, and secondly to evaluate their validity and accuracy in comparison with the 9th Version of the International Classification of Disease (ICD9) tabular listing published by the World Health Organization (WHO). As a result of this process some recommendations and suggestions on using the MSP ICD9 codes are put forward. The quality of the diagnostic codes in the PMEMASTER are generally satisfactory, though this quality varies between the different types of codes. In summary, the following conclusions have been drawn from this study:

1. All diagnostic codes in the PMEMASTER are left-justified with no blanks in front of the codes.
2. The major code types are mainly three, four, and five digit numeric codes, and codes containing the characters L, X, V, Z, A and B.
3. The diagnostic codes in the PMEMASTER are a mixture of a complete listing of ICD9 and a partial of listing of ICD9-CM codes, and therefore may be referred to as ICD codes rather than either ICD9 or ICD9-CM codes.
4. All standard three digit numeric codes and L, X, Z, A, and B codes are fully represented in the PMEMASTER; about 87 percent of all standard four digit codes; 70 percent of all standard five digit codes ; and 67 percent of V codes are found in the PMEMASTER. In total, approximatly 87 percent of all standard ICD9 codes are covered by the PMEMASTER.
5. Overall, more than 96 percent of patient counts, paid services, and paid amounts

in fiscal year 1994/95 are associated with those diagnostic codes in the PMEMASTER which match the WHO's standard ICD9 listing. Particularly, paid services and amounts associated with the matched three digit numeric codes, and codes with L, X, V and Z characters amount to more than 98 percent of the total services and amounts associated respectively with all these codes, while matched four digit codes contribute to about 88 percent of the total services and amounts affiliated with all four digit codes.

6. The accuracy and reliability of three digit numeric codes and codes with L, X, V and Z are more reliable than four digit numeric codes, while five digit numeric codes do not have a satisfactory level of accuracy or reliability.

Since most patient counts, paid services, and paid amounts in the PMEMASTER are associated with three digit numeric codes, the ICD codes in the PMEMASTER are not suitable for use with applications which specifically rely on four and five digit numeric diagnostic codes.

Table of Contents

I. Background

There have been increasing demands for information which is related to specific diagnoses in the area of ambulatory care services. The Medical Services Plan (MSP) of British Columbia has been requested to provide information on diagnostic codes and their related service utilization for different users, including epidemiologists, health care service researchers, and other professionals, etc. Recently, a pilot project of case-mix adjustment for patient classification, based on their potential utilization of health services, has been carried out in the Program Monitoring and Information Management Branch (PMIM), Resource Management Division, MSP. This project requires particularly specific and accurate diagnostic information for each patient, represented by both four and five digit ICD9 codes, (i.e. the WHO's 9th Version of International Classification of Diseases).

However, it has long been argued by users of MSP's diagnostic data that the quality of ICD9 information in the MSP claim database (PMEMASTER) is unreliable. Generally, there are two major problems related to the accuracy of the ICD9 codes. The first is that some physicians may not use an accurate code for patients, or may use inadequate codes for their diagnoses. For instance, diagnoses listed in WHO's ICD9 table as general symptoms (ICD9 780-789) , are commonly given by physicians as diagnoses for patients who may not present any diagnosable, or disease-specific symptoms, but who actually suffer from totally different diseases than these general symptom codes would indicate. This non-specific diagnosis therefore will categorize patients with different diseases into the same diagnostic group. Another problem is related to the MSP's claims information system which somehow mis-records ICD9 codes submitted by practitioners. A typical error is the decimal position problem. It has been found, for example, that codes 055.0 (measles with encephalitis) and 550 (inguinal hernia) were both reported as 0550 in the PMEMASTER.

It is not surprising to have such errors in a large database. The question is how serious are these problems and whether or not they impact the usefulness of the ICD9 related information from the PMEMASTER? Therefore, it is necessary to investigate and evaluate the quality of diagnostic information in the PMEMASTER in order to responsibly provide accurate ICD related information, and rationally make use of this information for various purposes.

To achieve this goal of accurate diagnostic codes, a project investigating the quality of diagnostic codes in the PMEMASTER has been undertaken. The purpose of the project is as follows:

1. investigating ICD9 codes in the PMEMASTER - what do they look like?
2. examining problems associated with the codes.
3. estimating both the magnitude of the problems and their impact on the use of the codes by researchers, etc.
4. proposing recommendations to improve and overcome the problems.

II. Data and Methodology

The analyses will be divided into several steps as described below. The data used in the analyses will be one of the PMEMASTER datasets, PUBLIC.PM9495, which contains all claims data for fiscal year 1994/95, and SASDATA.DIAGCODE which is a SAS format data set containing a full list of WHO’s official ICD9 codes.

Step One: *Background Analysis*

The PMEMASTER is the MSP database which contains a series of variables (data elements). These elements have been used in PMIM’s routine projects such as program monitoring and service utilization analyses. One of the elements, named DIAGCD, contains diagnostic codes, ie. ICD codes, and is originally designed to contain up to a 5 digit character variable. The first step of the analyses focuses on this element, including:

1. percentage distribution of ICD codes by of code types. The code types will be classified by the number of digits appearing in the variable DIAGCD.
2. percentage distribution of code frequency, distinct patient counts, paid services, and paid amounts by the code types.

It is hoped that the results will illustrate, in general terms, what the ICD codes in the PMEMASTER look like, and their associated patient counts, services, and paid amounts.

Step Two: *Comparison of DIAGCD and SASDATA.DIAGCODE*

The SASDATA.DIAGCODE is a SAS data set which contains a ***FULL*** listing of the WHO's ICD9 codes. Therefore, it is reasonable to use this data set as a standard tool to compare all the codes which are submitted by physicians and recorded in the PMEMASTER. Theoretically speaking, if a submitted code is correct, it will match one of those codes contained in the SASDATA.DIAGCODE. Any disparity between the submitted codes and the codes in the SASDATA.DIAGCODE will indicate that the submitted code is not an ICD9 code. Combining the results from the first step, rough estimates of the percentage of paid services and amounts associated with 'Correct' and 'Incorrect' codes will be derived. The detailed analyses in this step includes:

1. matching all ICD codes in the PUBLIC.PM9495 with the SASDATA.DIAGCODE.
2. analyzing percentage distribution of code frequency, patient counts, paid services and paid amounts by the code types within matched and unmatched code groups.
3. estimating the loss of information with regards to patient counts, paid services

and paid amounts if data is extracted based on the matched codes only.

Step Three: *Work on Unmatched Codes*

If the SASDATA.DIAGCODE contains a full list of WHO's ICD9 codes, the unmatched codes in the PMEMASTER will then be either incorrect, or belong to a different coding system such as ICD9-CM or ICD9-CM(Procdures). If this is not the case, the SASDATA.DIAGCODE might not contain the WHO's full tabular listing. In order to clarify these questions and to estimate the potential impact from these unmatched codes, a relatively detailed analyses will be performed on them in this step. The analyses will include:

1. percentage distribution of the unmatched codes and associated paid services and amounts by the code types(number of digits).
2. grouping fee items which are associated with these unmatched codes.
3. grouping practitioners who consistently submit a particular type of these unmatched codes.

Step Four: *Impact Estimation*

Based on the results from steps 1, 2, and 3, an estimate of the impact of incorrect ICD codes will be made and this will provide answers to the questions about what percentage of patient counts, paid services, and paid amounts would be underestimated if the diagnosis in the PMEMASTER is based on a three, four, or five digit ICD code.

Step Five: *MSP ICD Data Use Procedures*

A series of rules regarding how to use the MSP ICD data will be determined based on the previous analyses. It is expected that the impact of incorrect ICD information in the

PMEMASTER diagnostic codes will be minimized if users follow these newly determined rules in extracting and summarizing ICD data.

III. Results and Discussions

1. Diagnostic Codes in PMEMASTER - What do they look like?

DIAGCD, a 5 digit character field data element (variable) in the PMEMASTER, contains diagnostic codes which have been used as ICD9 codes within the PMIM Branch. The advantage of using a character variable is that it takes zero (0) as an independent digit with a meaningful value. For example, 0150 (Tuberculosis of Vertebral Column) in ICD9 is a different code from 150 (Malignant Neoplasm of Esophagus). The 0 is recognized as having a meaningful value. There are a number of large databases within the MOH, such as Hospital Programs and Vital Statistics, in which diagnostic codes are recorded in the same way. Also, many published papers present ICD9 codes in character form, such as 0150, instead of numeric form like 15.0. However, the recognition of 0 value in character variables also causes confusion in some circumstances. To illustrate, the original diagnostic code 150 may be mistakenly submitted as 0150 or 00150. If the DIAGCD were a numeric field, 0150, 00150, and 150 would be treated as the same code and 150 would be what appeared in the database. The errors made by such submissions would not impact on the accuracy of code 150 in its diagnostic meaning, however, the character variable of ICD9 would recognize these three codes as different diagnoses.

According to the Claims System of MSP, the structure of ICD9 codes in the PMEMASTER is left-justified five digit with the decimal concealed between the 3rd and 4th digit. Therefore, a ***CORRECT*** code submitted by physicians must have at least 3 digits with or without zero at the beginning and be left justified. If this is true, the appearance of ICD9 codes in the DIAGCD field submitted by physicians will look as follows:

Original Diagnostic Code		Appearance in the DIAGCD	
00X	- 3 digit code	00X__	- 3 digit, ended with 2 blanks.
00X.0	- 4 digit code	00X0_	- 4 digit, ended with 1 blank.
00X.00	- 5 digit code	00X00	- 5 digit, without blank.

Table 1 (Percent Distribution of Code Frequency and Associated Paid Service and Paid Amounts By Types of ICD Codes for Fiscal Year 1994/95) summarizes paid services and paid amounts by types of ICD codes appearing in the PMEMASTER, as well as frequencies and distinct counts for each code type. Figures 1.1 (Percent Distribution of ICD Code Frequency and Distinct ICD Code across Type of ICD Codes) and 1.2 (Percent Distribution of Paid Services and Paid Amounts across Type of ICD Codes) graphically display percentage distributions of code frequency, distinct code counts, and associated paid

Table 1: Percent Distribution of Code Frequency and Associated Paid Services and Paid Amounts By Types of ICD Codes for Fiscal Year 1994/95

Types of	ICD Code		Distinct ICD Code		Paid Services		Total Amounts	
ICD Codes	Frequency	Percent	Count	Percent	Services	Percent	Amounts	Percent
1. Numberic Codes								
Two Digits: 0X___	1	0.0	1	0.0	1	0.0	$58	0.0
XX___	34	0.0	24	0.2	44	0.0	$1,356	0.0
Subtotal	*35*	*0.0*	*25*	*0.2*	*45*	*0.0*	*$1,414*	*0.0*
Three Digits: 00X__	74,530	0.4	9	0.1	101,943	0.2	$2,809,246	0.2
0XX__	197,246	1.1	89	0.8	304,991	0.6	$8,885,843	0.7
XXX__	11,534,243	64.7	900	7.6	22,560,879	44.0	$789,702,783	58.7
Subtotal	*11,806,019*	*66.2*	*998*	*8.5*	*22,967,813*	*44.8*	*$801,397,872*	*59.6*
Four Digits: 00XX_	12,284	0.1	61	0.5	16,751	0.0	$474,222	0.0
0XXX_	48,569	0.3	592	5.0	93,295	0.2	$2,752,049	0.2
XXXX_	1,527,117	8.6	5,499	46.6	2,854,509	5.6	$107,833,879	8.0
Subtotal	*1,587,970*	*9.0*	*6,152*	*52.1*	*2,964,555*	*5.8*	*$111,060,150*	*8.2*
Five Digits: 00XXX	38,942	0.2	537	4.6	73,581	0.1	$2,947,241	0.2
0XXXX	2,984	0.0	176	1.5	5,333	0.0	$151,992	0.0
XXXXX	177,241	1.0	2,737	23.2	291,963	0.6	$10,668,976	0.8
Subtotal	*219,167*	*1.2*	*3,450*	*29.3*	*370,877*	*0.7*	*$13,768,209*	*1.0*
Total Numeric Codes	***13,613,191***	***76.4***	***10,625***	***90.1***	***26,303,290***	***51.3***	***$926,227,645***	***68.8***
2. Character Codes								
A or B Code	1,100,562	6.2	333	2.8	2,456,663	4.8	$66,733,654	5.0
L Code	1,662,670	9.3	31	0.3	16,894,371	33.0	$185,549,934	13.8
V Code	85,544	0.5	444	3.8	123,830	0.2	$3,699,091	0.3
X Code	1,094,523	6.1	11	0.1	2,191,034	4.3	$110,018,759	8.2
Z Code	244,157	1.4	23	0.2	3,230,393	6.3	$51,982,074	3.9
Others	33,379	0.2	330	2.8	50,962	0.1	$1,582,814	0.1
Total Character Codes	***4,220,835***	***23.7***	***1,172***	***9.9***	***24,947,253***	***48.7***	***$419,566,326***	***31.3***
All Type Codes	**17,834,026**	**100.0**	**11,797**	**100.0**	**51,250,543**	**100.0**	**$1,345,793,971**	**100.0**

services and paid amounts. Both the Table and Figures show different types of ICD codes, frequently used code types, and code types with a higher percentage of paid services and amounts.

Figure 1.1: Percent Distribution of ICD Code Frequency and Distinct ICD Code across Type of ICD Codes, B.C. 1994/95

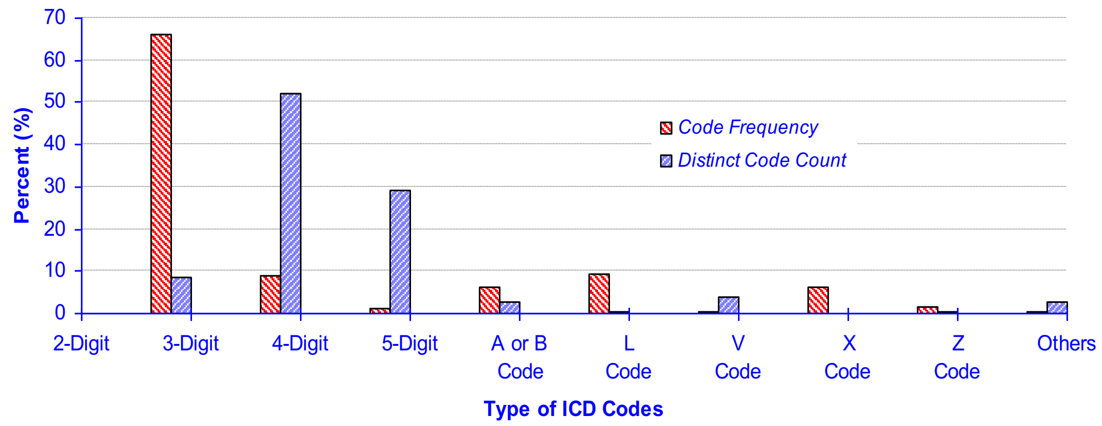
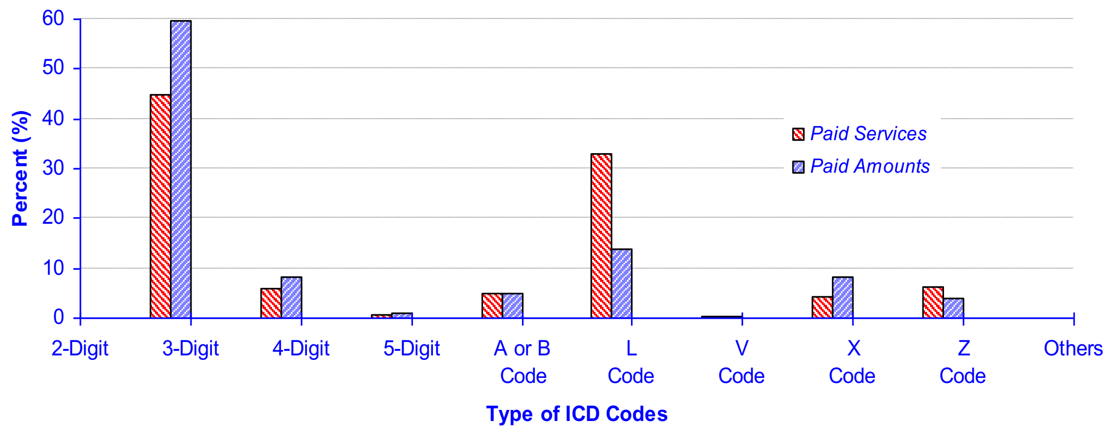


Figure 1.2: Percent Distribution of Paid Services and Paid Amounts across Type of ICD Codes, B.C. 1994/95



The ICD codes in the PMEMASTER are generally divided into two types, the first type being composed of numbers only, and the second type containing a mixture of characters and numbers. Within the numeric codes there are five sub-types of codes, based on the number of digits a code has. As an example 'X' indicates a numeric digit, from 0 to 9, thus code type 'XXXXX' means that this is a five digit numeric code and 'XXXX_' refers to four digit numeric code. In order to investigate the role of the digit zero, code sub-types within each code type are further divided, with an emphasis on the position of zero by the different code sub-type. For example, within the five digit numeric codes, 00XXX indicates that at least the first two digits are zeros; 0XXXX means that the first digit is zero but the second one is not zero; XXXXX refers to those five digit codes which do not have zeros in the first two digit positions. The same logic is also applied to two, three, and four digit numeric codes.

The codes with characters (character codes) are divided into seven sub-types, i.e. codes with A or B; with X; with L; with Z, with E; with V; and others. In WHO's official ICD9 listing, only codes starting with V and E are valid 'character' codes. In the PMEMASTER, however, several other character codes were created internally to represent particular services within the MSP, such as X for X-ray examinations, L for LAB tests, Z for anaesthetic, etc.

The purpose of dividing the diagnostic codes in the PMEMASTER into such detailed sub-groups is to investigate: 1) code validity, 2) most frequently used codes and their associated paid services and amounts, and 3) the potential impact of invalid codes. For instance, a valid numeric code must have at least three digits, therefore, two digit numeric codes (0X or XX) are invalid codes; the maximum number of digits in the numeric ICD9 codes is four, therefore, a five digit numeric code is not a valid ICD9 code.

The second column of Table 1 contains the frequency of different types of ICD codes and their percentage of the total number of codes used for one fiscal year (1994/95). In general, this column shows how frequently a code type is used by physicians. The most frequently used ICD9 codes in the PMEMASTER, for example, are three digit numeric codes (ie. XXX__, 0XX__ ,or 00X__) which accounts for more than 66 percent of all types of ICD codes recorded in the PMEMASTER. In addition, the paid services and paid amounts associated with this code type comprise approximately 45 and 60 percent of the totals, respectively, for fiscal year 1994/95. Following the three digit codes, other frequently used ICD codes are L (9.3 percent of all codes), accounting for 33 percent of total paid services and 14 percent of total paid amounts. The other codes represent, as a percentage of all codes, the following: four digit code s(9.0 percent), A and B codes (6.2 percent), and X code (6.1 percent). Five digit and the rest of the character codes were not frequently used codes (3.3 percent of all codes), for which the associated paid services and paid amounts consist of a small proportion of the totals (less than 8 percent of the total services and amounts associated with all codes for fiscal year 1994/95).

On further examinion of Table 1 and Figures 1.1 and 1.2, it is revealed that the percentage distribution of distinct ICD codes shows a different pattern from that of code frequency. The "distinct" here means "unique" ICD code. For example, if an ICD9 code 150 (malignant neoplasm of esophagus) was used twice, the frequency of this code would be two, but the distinct count of this code would be one. Total distinct ICD9 codes appearing in the PMEMASTER for fiscal year 1994/95 amount to 11,797 (Table 1), of which the four digit codes contributed over 50 percent, followed by five digit codes (30 percent) and three digit ones (8.5 percent).

The number of digits in an ICD codes represents the degree of diagnostic detail or specificity. In WHO's ICD9 Tabular Lists, three digit codes are major diagnostic group codes, while four digit codes give more specific diagnostic information. For example, code

150 represents malignant neoplasm of esophagus, a general term including a group of specific malignant neoplasms related to the esophagus, such as malignant neoplasms at cervical esophagus (1500), thoracic esophagus (1501), abdominal esophagus (1502), upper third of the esophagus (1503), middle third of the esophagus (1504), lower third of the esophagus (1505), etc. These specific diagnoses are indicated by the fourth digit. It is reasonable to assume that the details of diagnosis given by a physician in an office visit encounter (also known as ambulatory care service) will not only depend on general examination, but also, to a large extent, the evidence from a series of associated tests such as Lab tests, X-ray examinations, and so on. Therefore, it is expected that most of the diagnoses given by family physicians at the time of an office visit will be more general rather than specific because of the lack of enough clinical and pathological evidence for detailed and specific diagnoses. This means that most of the ICD9 codes used by physicians for office visits will be three digit (group diagnoses) rather than four digit (specific diagnoses). Our data confirmed the hypothesis that the most frequently used diagnostic codes (codes with higher frequency, see Table 1 and Figure 1.1) are the three digit ones and are associated with the largest proportion of paid services and amounts.

A conlusion is that the pattern of the ICD code types recorded in the PMEMASTER reflects the characteristics of ambulatory care services in which three digit codes are the most frequently used.

An unusual finding from this analysis is that there is a small proportion of five digit numeric codes existing in the PMEMASTER. As we know, there are no five digit numeric ICD9 codes in the WHO's official ICD9 tabular listing. The potential explanations for their existence would be that either they are not valid codes (mistakes made by physicians) or they belong to another coding system for diagnostic classification, such as ICD9-CM (International Classification of Disease, 9th Version, Clinical Modification). If the latter is true, the ICD codes in the PMEMASTER are a mixture of different coding systems.

2. Diagnostic Codes in PMEMASTER - Are they valid?

As shown in Table 1 and Figures 1.1 and 1.2, the most frequently used types of diagnostic codes in the PMEMASTER are the three digit numeric and L codes, which account for more than 75 percent of total codes recorded in the data, and over 77 percent of total paid services in one fiscal year. If these codes are valid (based on whether they exist in the official ICD9 listing or not), then we can, with a certain amount of confidence, fullfil diagnostic-code-specific service and expenditure requests by users who want to use the MSP diagnostic information. In order to confirm the validity of these codes, a computerized version of the official ICD9 listing is needed to match all existing ICD codes in the PMEMASTER. Fortunately, a SAS data set, named SASDATA.DIAGCODE, contains a full list of WHO's official ICD9 codes. Table 2.1 (Percent Distribution of Distinct ICD Codes in....) presents percentage distribution of the different code types in the SASDATA.DIAGCODE. There are a total of 7,137 distinct ICD codes contained in the SASDATA.DIAGCODE, of which the four digit numeric codes account for over 72 percent, followed by the 3 digit numeric codes (12.8 percent) and those with the character V (8.6 percent). It is surprising to find, however, that there are a few of the five digit numeric codes in this list as they should not exist in an official ICD9 list. Therefore, the codes in the SASDATA.DIAGCODE are probably inter-mixed with other coding systems, such as ICD9-CM, which contains five digit numeric codes.

Table 2.1: Percent Distribution of Distinct ICD Codes in SASDATA.DIAGCODE*, for Fiscal Year 1994/95

Type of Code	Distinct Count	Percent
3 Digit Numeric Code (XXX__)	912	12.8
4 Digit Numeric Code (XXXX_)	5,176	72.5
5 Digit Numeric Code (XXXXX)	359	5.0
Code with Z Character (Anaesthetic)	1	0.0
Code with L Character (LAB Tests)	1	0.0
Code with A or B Characters	52	0.7
Code with V Character	615	8.6
Code with X Character (X-ray)	1	0.0
Other Character Codes	20	0.3
All Types	***7,137***	***100.0***

** It is a SAS format dataset containing a list of WHO's official ICD9 codes.*

Table 2.2 (Matches between All Distinct ICD Codes in PMEMASTER...) is the comparison between all distinct ICD codes contained in the SASDATA.DIAGCODE (hereafter referred to as the *standard*) and those contained in the PMEMASTER (for all claims submitted for fiscal year 1994/95). The first column in the Table presents the code types, the second and third columns show the number of distinct codes in the PMEMASTER that match to the *standard* list and associated paid services. Columns four and five contain the same information for those codes which do not match to the *standard* and the last two columns are the summary for all codes in the PMEMASTER for 1994/95.

Table 2.2: Matches between All Distinct ICD Codes in PMEMASTER and SASDATA.DIAGCODE*
Fiscal Year 1994/95

Types of ICD Codes	Codes Matched to DIAGCODE				Codes Not Matched to DIAGCODE				All Distinct ICD Codes			
In PMEMASTER	*Count*	*%*	*Services*	*%*	*Count*	*%*	*Services*	*%*	*Count*	*%*	*Services*	*%*
2 Digits (XX___)	.	.	.	.	25	100.0	45	100.0	25	100.0	45	100.0
3 Digits (XXX__)	912	91.4	22,957,909	100.0	86	8.6	9,904	0.0	998	100.0	22,967,813	100.0
4 Digits (XXXX_)	4,521	73.5	2,634,416	88.9	1,631	26.5	330,139	11.1	6,152	100.0	2,964,555	100.0
5 Digits (XXXXX)	249	7.2	30,030	8.1	3,201	92.8	340,847	91.9	3,450	100.0	370,877	100.0
Total Numeric Codes	*5,682*	*53.5*	*25,622,355*	*97.4*	*4,943*	*46.5*	*680,935*	*2.6*	*10,625*	*100.0*	*26,303,290*	*100.0*
V Code	412	92.8	121,948	98.5	32	7.2	1,882	1.5	444	100.0	123,830	100.0
L Code (LAB)	1	3.2	16,744,714	99.1	30	96.8	149,657	0.9	31	100.0	16,894,371	100.0
X Code (XRay)	1	9.1	1,891,665	86.3	10	90.9	299,369	13.7	11	100.0	2,191,034	100.0
Z Code (Anaethesis)	1	4.3	3,202,355	99.1	22	95.7	28,038	0.9	23	100.0	3,230,393	100.0
A & B Codes	52	15.6	2,300,817	93.7	281	84.4	155,846	6.3	333	100.0	2,456,663	100.0
Other Character Code	7	2.1	42,064	82.5	323	97.9	8,898	17.5	330	100.0	50,962	100.0
Total Character Codes	*474*	*40.4*	*24,303,563*	*97.4*	*698*	*59.6*	*643,690*	*2.6*	*1,172*	*100.0*	*24,947,253*	*100.0*
All Types	**6,156**	**52.2**	**49,925,918**	**97.4**	**5,641**	**47.8**	**1,324,625**	**2.6**	**11,797**	**100.0**	**51,250,543**	**100.0**

**** A SAS format dataset containing a full list of official WHO's ICD9 codes.***

For the two digit code type, none of the total 25 distinct codes match to the *standard*, meaning they do not exist in the *standard*. As for the three digit code type , 912 out of a total of 998 distinct codes (91.4percent) are associated with nearly 100 percent of the paid services and match to the *standard*; 73.5 percent of four digit codes in association with 88.9 percent of services match to the *standard*, while only 7.2 percent of five digit codes with 8.1 percent of total services match to the *standard*. In summary, 53.5 percent of all distinct numeric codes match to the *standard*, which contribute to 97.4 percent of total services.

Among character codes, 412 out of 444 V-codes (92.6 percent) match to the *standard*, with 98.5 percent having associated paid services. While the rest of the character codes do not match to the *standard* very well (percent of match ranging from 2.1 percent to 15.6 percent), the paid services associated with the matched codes, however, have a very high percentage match to the total services, ranging from 82.5 percent (other character codes) to 99.1 percent (L-code). In general, about 40.4 percent of all character codes match to the *standard*, but have a high association of 97.4 percent of total paid services.

In summary, there are a total of 11,797 distinct diagnostic codes in the PMEMASTER for fiscal year 1994/95, of which 6,156 (about 52.5 percent) match to the codes contained in the SASDATA.DIAGCODE (total 7,137) which is supposed to be a complete listing of the WHO's official ICD9 codes. The match rate varies by different code types, with the highest rate for V-codes (92.8 percent), followed by three digit numeric codes (91.4 percent), and four digit numeric codes (73.5 percent). Table 2.3 (Comparison of Distinct ICD codes between SASDATA.DIAGCODE and...) shows the percentage of the *standard* codes covered by the PMEMASTER across different code types. In total, 86.3 percent (6,156 out of 7,137) of standard ICD9 codes are covered in the claims data, PMEMASTER. The three digit numeric codes, and the character codes L, Z, and X are 100 percent covered, while 87.3 percent for four digit numeric codes, and 69.4 percent for five

Table 2.3: Comparison of Distinct ICD codes between SASDATA.DIAGCODE* and PMEMASTER for Fiscal Year 1994/95

Type of Code	Distinct ICD9 in DIAGCODE	Matched Distinct ICD9 in PMEMASTER	Matched Codes Percent	Associated PHN** Percent	Associated Service** Percent	Associated Amounts** Percent
3 Digit (XXX__)	912	912	100.0	99.9	100.0	100.0
4 Digit (XXXX_)	5,176	4,521	87.3	91.0	88.9	88.1
5 Digit (XXXXX)	359	249	69.4	7.4	8.1	10.9
Z Code(Anaesthetic)	1	1	100.0	98.7	99.1	99.1
L Code(LAB Tests)	1	1	100.0	98.0	99.1	98.9
A & B Code	52	52	100.0	93.5	93.7	92.2
V Code	615	412	67.0	98.7	98.5	99.3
X Code (X-ray)	1	1	100.0	83.1	86.3	88.2
Other Character Codes	20	7	35.0	80.0	82.5	70.9
All Types	***7,137***	***6,156***	***86.3***	***96.3***	***97.4***	***96.5***

** It is a SAS format dataset containing a full list of WHO's official ICD9 codes.*

*** PHNs, Services, Amounts are associated with those matched ICD9 codes in the PMEMASTER.*

digit numeric codes. It is reasonable, therefore, to conclude that the accuracy of those matched codes is satisfied because they cover more than 86 percent of standard ICD9 codes. The assoication of service utilization with diagnostic codes (measured by paid services and paid amounts) is even better because these matched codes are associated with approximate 97 percent of total paid services and total paid amounts for one fiscal year. Total patient counts associated with these matched codes cover about 97 percent of all patients.

3. Using the Diagnostic Codes in PMEMASTER

To confirm the findings outlined above, it is useful to use some practical examples to analyze the quality of the ICD codes in the PMEMASTER. Here are three projects related to diagnostic code information which have been recently conducted in PMIM.

Project One: Comparison with Manitoba

Manitoba reports service utilization by major ICD9 groups (three digit codes) every year through their annual report. In order to evaluate the accuracy of the ICD codes in the PMEMASTER, the percentage distribution of paid services and paid amounts by major ICD9 groups for British Columbia in fiscal year 1993/94 were compared to Manitoba's 1993/94 annual report. Figures 3.1 (Percent Distribution of Paid Services Across Major ICD9...) and 3.2 (Percent Distribution of Paid Amounts across Major ICD9...) show this comparison.

The major groups were simply created by taking the first three digits from each individual ICD code in the PMEMASTER. The logic behind this truncating is that all the ICD codes in the PMEMASTER are left-justified, and all four and five digit codes are sub-divisions of the three digits, and thus can be converted to three digit codes by simply using the first three digits of actual code.

Figure 3.1: Percent Distribution of Paid Services across Major ICD9 Categories, Comparison between B.C. & Manitoba, 1993/94

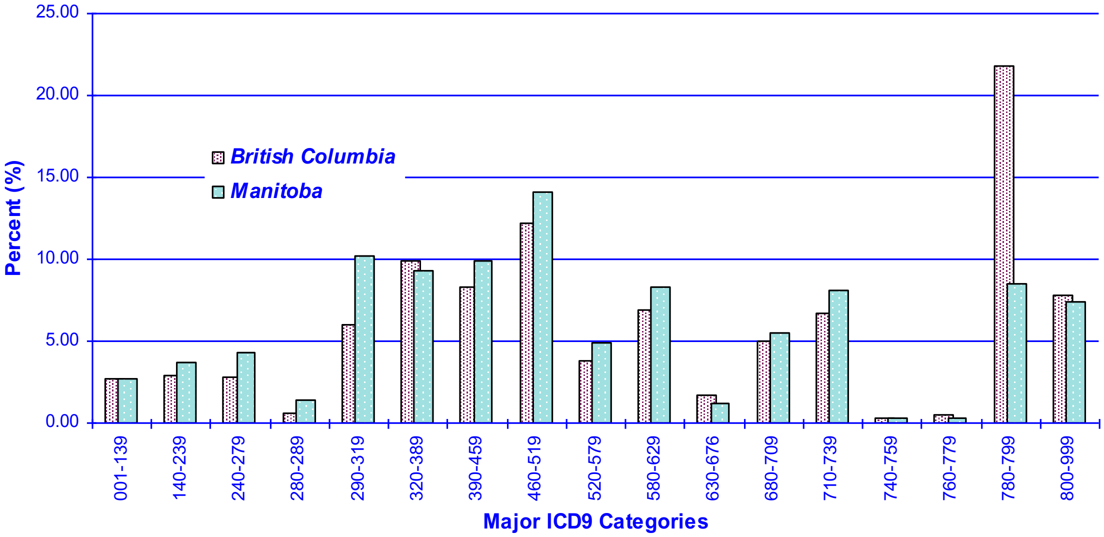
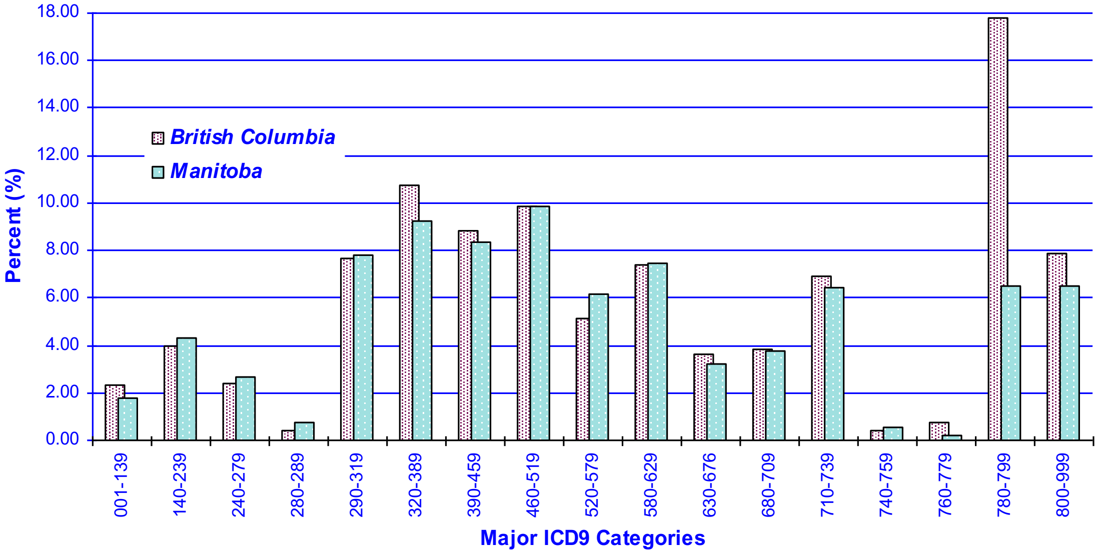


Figure 3.2: Percent Distribution of Paid Amounts across Major ICD9 Categories, Comparison between B.C. & Manitoba, 1993/94



The two figures presented a similar pattern in terms of percent distribution of paid services and paid amounts across these major ICD groups between the two provinces, although they are not exactly the same. The interpretation of this similarity is that the service utilization generated by these major diagnoses codes is comparable between these two provinces. In other words, the accuracy of diagnostic information in the PMEMASTER is reliable with the assumption that Manitoba's data is accurate (based on the fact that their data has been published).

The major difference, however, between these two provinces is that the percentage of total paid services and amounts contributed by codes 780-799 for B.C. data are twice those of Manitoba's. Further analyses of this group of codes (the results are not shown) found that there are two major diagnoses categories covered by these codes, i.e. general symptoms (780-789) and nonspecific abnormal findings (790-796). But more than 95% of the contribution to total services and amounts related to this diagnostic group (780 - 799) are from codes 780-789 (general symptoms). It seems, therefore, B.C. physicians are more likely to use this group of ICD codes as a dianoses for these non-specific symptoms.

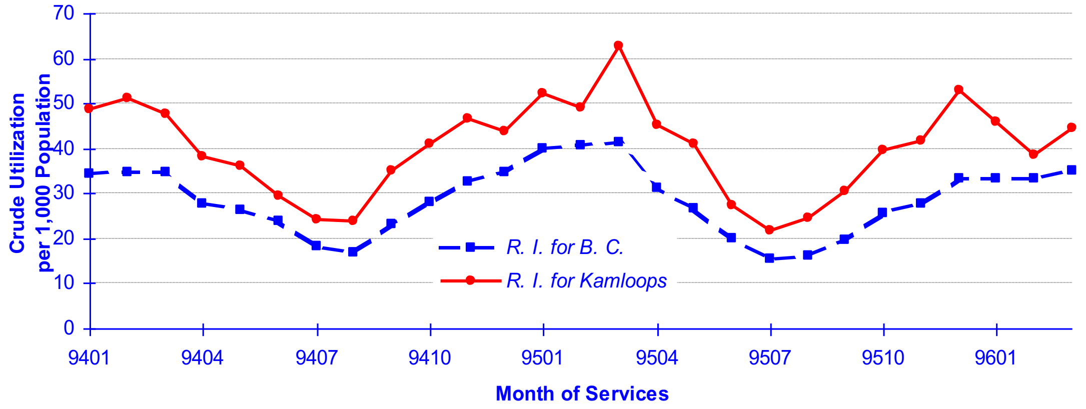
In summation, this practical example confirms, to some extent, the relative accuracy of the diagnostic information represented by the ICD codes in the PMEMASTER The patterns of percentage distributions of service utilization and paid amounts across major diagnostic groups (represented by the three digit codes) are comparable to Manitoba's.

Project Two: Kamloops' Project

Another recent project related to diagnostic information is the Kamloops' Project. The background on this project is the concern last spring about the turbidity of the drinking water in the city of Kamloops being far higher than the Canadian Drinking Water Guide recommended level. The Medical Health Officer suspected that this high turbidity may have caused an endemic outbreak of Gastrointestinal (GI) problems which then caused higher utilization in related GP office visits. He requested a time series analysis of service utilization for GI related office visits with diagnoses represented by ICD9 codes 001-009, 558, and 787. He also suggested that service utilization for Respiratory Infections (RI), ICD9 codes 460-466, be the control group for comparison.

Figure 3.3 (Services Utilization for GP Visits for Respiratory Infections...) presents the time trend of service utilization for respiratory infections on a monthly basis from January 1994 to March 1996 for both Kamloops (Local Health Area 24) and the Province as a whole. There are distinguishable seasonal trends for both Kamloops and the province where the utilization rates rise in the months from late fall through winter to early spring, and drop in the summer months. Logically, this seasonal trend is reasonable because the colder seasons may have a greater affect on people suffering from respiratory infections (including bacteria and virus infections), thus causing a higher utilization rate for R.I. services. The consistent higher utilization rate in Kamloops, in comparison to the Province as a whole, is probably due to the colder winter season in Kamloops. This example shows that the MSP ICD codes capture satisfactory diagnostic information, and therefore indirectly demonstrates that the

Figure 3.3: Services Utilization for GP Visits for Respiratory Infections (R.I.)*
For Kamloops and the Province, by Month of Services, January 1994 to March 1996



** ICD9 Codes: 460-466.*

diagnostic information from the PMEMASTER based on ICD codes are relatively accurate and reliable.

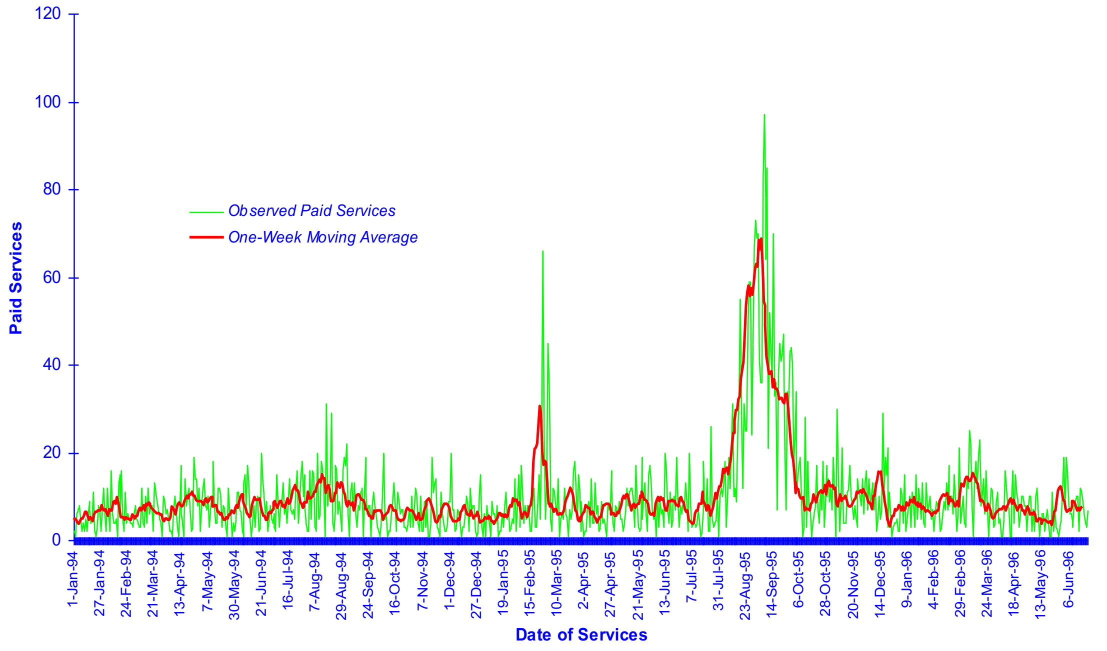
Project Three: Revelstoke's Outbreak

In August of 1995, there was an outbreak of a water-borne disease in Revelstoke (Local Health Area 19), British Columbia. The major symptoms were gastrointestinal disorders. According to the theory of Epidemiology, there is an increase of patients with similar symptoms occurring before the first specific diagnosis. In other words, an abnormal increase of patients with similar symptoms in an area may trigger local physicians' attention, causing them to investigate the possibility of a disease outbreak. Therefore, it is expected that an abnormally high service utilization of GPs' office visits will appear in Revelstoke, at the time of the disease outbreak. Figure 3.4 (Paid Services for GPs' Office Visits for Gastrointestinal Illness ...) shows a time series analysis of daily service utilization of GPs' office visits (with a one-week moving average smoothing[1]) for Revelstoke from January of 1994 to June of 1996. The services were extracted from the PMEMASTER, based on a list of three digit diagnostic codes which are related to gastrointestinal illness.

It can be seen that the daily service utilization for Revelstoke has been stable for the last two years, except the time period from the end of July of 1995 to early October of 1995 where there was a significant rise in daily utilization of GI types of services. This pattern is in accordance with the reported local outbreak of water-borne disease in Revelstoke. On the other hand, this analysis has illustrated that the ICD codes in the PMEMASTER do capture, timely and accurately, diagnostic information from GPs' services.

[1] One -week moving average is used here to reduce daily fluctuation in service utilization. This is a common statistical method employed in time series analysis. A more detailed explanation can be found in statistical texts.

Figure 3.4: Paid Services for GPs' Office Visits for Gastrointestinal Illness (G. I.)
Revelstoke (LHA 19), January 1, 1994 to June 30, 1996



In summary, the investigations previously outlined demonstrated that most of the diagnostic codes in the PMEMASTER are reliable and accurate. This conclusion has been further confirmed by some practical examples which used selected ICD9 codes to query information for particular studies. The codes *per se* have also been extensively examined in terms of their validity. These results integrated together can lead to the following specific conclusions about the ICD codes in the PMEMASTER.

1. All ICD codes are left-justified with no blanks in front of the codes.
2. The major code types are numeric (mainly 3, 4, and 5 digits), and character codes (mainly L,X,V,Z,A and B).
3. All standard WHO's three digit numeric codes and L, X, V,Z, A and B character codes are included in the PMEMASTER; about 87% of all standard four digit, 70% of all standard five digit, and 67% of all standard V codes are covered in the PMEMASTER. Overall, about 87% of all standard ICD9 codes are contained

in the PMEMASTER.

4. Overall, more than 96% of paid services and paid amounts are associated with those ICD codes in the PMEMASTER which match to the standard list. Particularly, those matched three digit numeric codes and Z, L and V codes contribute to more than 98% of total services and paid amounts which are associated with these codes respectively while the matched four digit codes account for about 88% of total services and amounts associated with all four digit numeric codes (i.e. matched and unmatched four digit numeric codes).
5. The accuracy and reliability of diagnostic codes in the PMEMASTER are generally satisfactory.

4. Invalid Diagnostic Codes in the PMEMASTER - What do they look like?

The results previously outlined revealed that quite a few of the diagnostic codes (5,641 out of 11,797, about 48 percent, see Table 2.2) recorded in the PMEMASTER do not match to the standard ICD9 list. Although the paid services, amounts and patients associated with them are insignificant (less than 5 percent, see Table 2.3), it is still worthwhile investigating the potential reasons behind these unmatched codes.

4.1 Diagnostic Codes in the PMEMASTER: ICD9 or ICD9-CM?

One of the unexpected results from the analyses of code types contained both in the PMEMASTER and SASDATA.DIAGCODE is that there are a certain number of five digit numeric codes, which are invalid in WHO's official ICD9 list. However, another coding schedule, entitled ICD9-CM (International Classification of Diseases, 9th Version, Clinical Modification) contains valid five digit numeric codes. The ICD9-CM, a modified ICD9, was originally created in the United States based on WHO's ICD9, and published by the National Center for Health Statistics (NCHS) for American use. The intent of this modification is to use the ICD9-CM as a tool in the area of classification of morbidity data for indexing of medical records; medical care reviews; ambulatory and other medical care programs; and for basic health statistics. To describe the clinical picture of the patient, the diagnostic codes must be more precise than those the WHO requires (their major concern being for statistical groupings and trend analysis). As a result, the disease classification in the ICD9-CM has been expanded to include health-related conditions and provide a greater specificity at the fifth digit level of detail. These fifth digits are not optional; they are intended for use in recording the information substantiated in the clinical record. The modification of ICD9 to ICD9-CM can be summarized as follows:

1) Three-digit codes:
    - their categories, contents, and sequence are unchanged from ICD9,
    - no new three-digit codes are created,
    - those three-digit ICD9 codes, not having a fourth digit, are subdivided to four digits, where necessary, to add clinical detail and isolate terms for clinical accuracy,
2) Four-digit codes:
    - an extra one digit is added to MOST of the existing four-digit rubrics to create five-digit codes in order to achieve the desired detail,
    - a few of the new four digit codes unique to ICD9-CM were created based on the existing ICD9 three-digit codes when the necessary detail could not be accommodated by the addition of a fifth digit subclassification on the existing four digit code. These new four digit codes are marked with a special symbol. There are a total of 28 three-digit codes from which new four-digit codes were created.

Therefore, the major difference between ICD9 and ICD9-CM is that the former does not have five-digit numeric codes, but the latter does. In addition, ICD9-CM has a few four-digit numeric codes which do not exist in ICD9. However, all ICD9-CM codes can be collapsed back to ICD9. Therefore the conclusion drawn is that the diagnostic codes contained in both SASDATA.DIAGCODE and PMEMASTER may be ICD9-CM, or at least a mixture of ICD9 and ICD9-CM.

Further examination of those diagnostic codes unmatched in the PMEMASTER found that a certain number of these codes are valid ICD9-CM codes. Table 4.1(Segment of Diagnostic Codes in the PMEMASTER Which Do Not Match...) lists a segment of these unmatched codes.

Table 4.1: Segment of Diagnostic Codes in the PMEMASTER Which Do Not Match to SASDATA.DIAGCODE, Fiscal Year 1994/95

Unmatched but Valid ICD9-CM Codes	Unmatched but Unique ICD9-CM Four Digit Codes	Unmatched and Invalid ICD9 and ICD9-CM Codes
00320	2508	00100
00321	3124	00107
00322	4419	00110
00324	6442	00112
00841	8009	00117
00842	8016	00118
00849	8046	0012
01000	8047	00122
01100	8138	00125
01101	8238	0013
01140	8239	00132
0132	8518	00133
0133	8523	00135
0136	9415	0014
......	......	......

In summary, the following findings were found for those unmatched codes:

1) Approximately 83% of all unmatched codes (4,682 out of 5,641) are not in the listings of either ICD9 and ICD9-CM.
2) Approximately 16% of all unmatched codes (902 out of 5,641) are valid ICD9-CM codes.
3) Approximately 0.2% of all unmatched codes (12 out of 5,641) are unique ICD9-CM four digits codes (newly created four digits codes in ICD9-CM, but not

existing in ICD9).

These findings further indicate that the diagnostic codes in the PMEMASTER are ICD9-CM codes, or at least include some of ICD9-CM codes; and SASDATA.DIAGCODE seems to consist of both ICD9 and an incomplete tabular listing of ICD9-CM codes.

4.2 Unmatched Diagnostic Codes and Fee Items

The previous analyses showed that there are a considerable number of invalid diagnostic codes in the PMEMASTER (4,682 out of 11,797). It seems to be inappropriate to totally attribute these codes to physicians' mis-codings because they are so many. It is worthwhile, however, further investigating to determine if there is any pattern behind these codes. One action is to analyze the association between these unmatched codes and fee items because some fee items may generally represent particular services which are then related to particular diagnoses.

Table 4.2 (Percent Distribution of Paid Services and Paid Amounts by Types of Unmatched ICD Codes...) presents the percentage distribution of the frequency of unmatched distinct diagnostic codes and associated paid services and paid amounts. Among all unmatched codes, five and four digit numeric codes and X, L, and A & B codes contribute the largest proportion of paid services and amounts.

Table 4.2: Percent Distribution of Paid Services and Paid Amounts by Types of Unmatched ICD Codes In the PMEMASTER for Fiscal Year 1994/95

Types of Unmatched	Distinct ICD Code		Paid Services		Total Amounts	
ICD Codes	Count	Percent	Services	Percent	Amounts	Percent
1. Numberic Codes						
Two Digits: 0X___	1	0.0	1	0.0	$58	0.0
XX___	24	0.4	44	0.0	$1,356	0.0
Subtotal	*25*	*0.4*	*45*	*0.0*	*$1,414*	*0.0*
Three Digits: 00X__	0	0.0	0	0.0	$0	0.0
0XX__	8	0.1	1,820	0.1	$72,215	0.2
XXX__	78	1.4	8,084	0.6	$212,956	0.5
Subtotal	*86*	*1.5*	*9,904*	*0.7*	*$285,171*	*0.7*
Four Digits: 00XX_	17	0.3	1,313	0.1	$13,774	0.0
0XXX_	296	5.2	18,007	1.4	$636,721	1.4
XXXX_	1,318	23.4	310,819	23.5	$12,531,178	26.7
Subtotal	*1,631*	*28.9*	*330,139*	*25.0*	*$13,181,673*	*28.1*
Five Digits: 00XXX	537	9.5	73,581	19.8	$2,947,241	6.3
0XXXX	170	3.0	4,896	0.4	$139,200	0.3
XXXXX	2,494	44.2	262,370	5.6	$9,176,317	19.5
Subtotal	*3,201*	*56.7*	*340,847*	*25.8*	*$12,262,758*	*26.1*
Total Numeric Codes	***4,943***	***87.6***	***680,935***	***51.5***	***$25,731,016***	***54.9***
2. Character Codes						
A or B Code	281	5.0	155,846	11.8	$5,225,072	11.1
L Code	30	0.5	149,657	11.3	$2,064,922	4.4
V Code	32	0.6	1,882	0.1	$25,937	0.1
X Code	10	0.2	299,369	22.6	$12,976,532	27.6
Z Code	22	0.4	28,038	2.1	$453,817	1.0
Others	323	5.7	8,898	0.7	$460,813	1.0
Total Character Codes	***698***	***12.4***	***643,690***	***48.6***	***$21,207,093***	***45.2***
All Type Codes	**5,641**	**100.0**	**1,324,625**	**100.1**	**$46,938,109**	**100.0**

To investigate the association between these codes and fee items a selection was made of the type of five digit numeric codes, starting with two zeros (00XXX). The reason in selecting this code type is that it contributes a relatively high proportion of paid services and code frequency (19.8% and 9.5%, see Table 4.2); and it would belong to the first major ICD9 group (001-139: Infections and Parasitic diseases) if treated as either a valid ICD9 or ICD9-CM because the first three digits of these codes are between 001 and 009. Fee items associated with these codes should indicate some specific service related to these particular types of disease.

Due to large number of this code type (537 in total), a repeated random sampling method was used to extract a small number of codes and associated fee item codes, plus their paid services and paid amount. Table 4.3 (Association between Fee Items and Selected Unmatched ICD Codes) shows the results. The first column listed the number of samples (total 5), column 2 lists the associated fee item codes; column 3 contains original diagnostic codes recorded in the PMEMASTER; columns 4 to 6 list new codes derived by different extraction methods from the original codes, (i.e. first three, second three, and last three digits). Obviously, the codes in columns 4 and 5 will be in the same major ICD9 groups (001-139), while the codes in column 6 will be in different major ICD9 group, most of which are related to diseases of the nervous system and the sensory organs. For example, code 360 represents "Disorders *of the Eye Globe*"; code 361, "Retinal *Detachments and Defects*"; code 362, "Other *Retinal Disorders*"; and code 376, "Disorders *of the Orbit*", etc.

Checking the fee item codes, it is found that most of these fee items represent services related to eyes, and only those codes in column 6 are logically connected to them. It seems, therefore, that using the last three digits of this type of diagnostic code would represent a reasonable diagnoses. In other words, some of those invalid codes could be converted to meaningful ones by selectively extracting certain digits.

Table 4.3: Association between Fee Items and Selected Unmatched ICD Codes

Sample No.	Fee Item Code	Original ICD Code	1st 3 Digit Code	2nd 3 Digit Code	Last 3 Digit Code	Paid Service	Paid Amount
1	2038	00371	003	037	371	253	1,513
1	2046	00376	003	037	376	3	21
1	2116	00362	003	036	362	3	1,509
1	2120	00377	003	037	377	1	10
1	2164	00378	003	037	378	4	1,750
1	2171	00371	003	037	371	4	453
1	2174	00370	003	037	370	3	748
2	2010	00363	003	036	363	45	2,844
2	2010	00373	003	037	373	181	11,439
2	2011	00366	003	036	366	52	2,032
2	2015	00372	003	037	372	33	1,426
2	2018	00364	003	036	364	69	1,315
2	2048	00361	003	036	361	1	11
2	2049	00363	003	036	363	1	28
2	2049	00379	003	037	379	2	28
3	2018	00376	003	037	376	73	777
3	2019	00363	003	036	363	70	1,225
3	2040	00376	003	037	376	1	112
3	2043	00375	003	037	375	1	54
4	2017	00376	003	037	376	29	535
4	2019	00363	003	036	363	70	1,225
4	2040	00376	003	037	376	1	112
4	2056	00371	003	037	371	22	1,474
5	2011	00362	003	036	362	21	820
5	2011	00368	003	036	368	3	117
5	2015	00360	003	036	360	5	216
5	2015	00784	007	078	784	2	86
5	2046	00363	003	036	363	26	142
5	2048	00365	003	036	365	1	6

Based on this assumption, further analyses were performed on other code types. The following interesting findings were found:

1) L-Code (valid code: 01L):
   Ninety-nine percent of paid services and paid amounts were associated with only two types of unmatched L-codes, 01L0 and 01L00. When compared to the valid L-code, the first three digits of 01L0 and 01L00 would be reasonable to extract in order to convert them into a valid L code.
2) X-Code (valid code: 01X):
   Ninety-nine percent of paid services and paid amounts were associated with only two types of unmatched X-codes, 01X0 and 01X00. When compared to the valid X-code, the first three digits of 01X0 and 01X00 are again good extraction choices to convert them into a valid X code.
3) Z-Code (valid code: 01Z):
   Approximately 95 percent of paid services and paid amounts were associated with only two types of unmatched Z-codes, 01Z0 and 01Z00. When compared to the valid Z-code, the first three digits of 01Z0 and 01Z00 will be once again better extraction choices to convert them into valid Z code.
4) Five Digit Numeric Codes:
   Approximately 75 percent of paid services and 70 percent of paid amounts were associated with only two types of unmatched five digit numeric codes, 00XXX (two zeros as the first two digits) and XXX00 (two zeros as the last two digits). There is still a large proportion of these codes which cannot be reasonably converted. However, these unmatched five digits codes are based on the match to the codes in the SASDATA.DIAGCODE which, as stated previously, may not contain a full listing of valid five digit ICD9-CM codes. In other words, there are

probably a certain proportion of these remaining five digit codes which are actually valid ICD9-CM codes. To confirm this, a full listing of ICD9-CM code is needed.

5) Four Digit Numeric Codes:
There seems to be no systematic way to convert unmatched four digit numeric codes to valid ones. This is not unexpected based on the findings previously stated. These unmatched codes result from the match between the PMEMASTER and SASDATA.DIAGCODE, and the latter does not contain a full listing of either the ICD9 or ICD9-CM list. Therefore, there are probably a number of unmatched four digit codes which are actually valid ICD9 or ICD9-CM codes. To confirm this, a full listing of ICD9 or ICD9-CM codes is needed.
6) A and B Codes (valid codes: 01A, 01B):
Among all unmatched codes with characters A and B, extracting the first three digits as ‘01A’ and ‘01B’ did not improve matches significantly. For example, paid services and paid amounts associated with these converted A and B codes amount to only 2-3 percent.

In summary, the unmatched codes are probably attributable to two causes. The first is that the standard code listing (SASDATA.DIAGCODE) is incomplete, thus leaving some valid ICD9 or ICD9-CM codes unmatched. This has been confirmed by an eyeball check of the individual unmatched codes (see table 4.1). As for the second, a large proportion of codes have extra zeros in the first or last two digit positions which become valid by simply removing these zeros. This applies in particular to the five digit numeric codes and the character codes L, X, and Z. Table 4.4 (Match between All Distinct ICD Codes in PMEMASTER AND SASDATA.DIAGCODE...) shows the increase of services after first and second matches between the PMEMASTER and the SASDATA.DIAGCODE. Here, the first match is the match between original diagnostic codes in the PMEMASTER and the SASDATA.DIAGCODE; the second one is the match between the unmatched codes (for which the extra zeros have been removed), and SASDATA.DIAGCODE.

Table 4.4: Matches between All Distinct ICD Codes in PMEMASTER and SASDATA.DIAGCODE*
After First and Second Matches, Fiscal Year 1994/95

Types of ICD Codes	**1st Match****		**2nd Match****		**All Matched Codes**	
In PMEMASTER	***Services***	***%***	***Services***	***%***	***Services***	***%***
2 Digits (XX___)	-		-	-	-	-
3 Digits (XXX__)	22,957,909	100.0	-	-	22,957,909	100.0
4 Digits (XXXX_)	2,634,416	88.9	-	-	2,634,416	88.9
5 Digits (XXXXX)	30,030	8.1	255,797	69.0	285,827	77.1
Total Numeric Codes	*25,622,355*	*97.4*	*255,797*	*1.0*	*25,878,152*	*98.4*
V Code	121,948	98.5	-	-	121,948	98.5
L Code (LAB)	16,744,714	99.1	149,582	0.9	16,894,296	100.0
X Code (XRay)	1,891,665	86.3	298,795	13.6	2,190,460	100.0
Z Code (Anaethesis)	3,202,355	99.1	26,803	0.8	3,229,158	100.0
A & B Codes	2,300,817	93.7	-	-	2,300,817	93.7
Other Character Code	42,064	82.5	-	-	42,064	82.5
Total Character Codes	*24,303,563*	*97.4*	*477,062*	*1.9*	*24,780,625*	*99.3*
All Types	**49,925,918**	**97.4**	**732,859**	**1.4**	**50,658,777**	**98.8**

**** A SAS format dataset containing a full list of official WHO's ICD9 codes.***

***** First match is between original codes in the PMEMASTER and SASDATA.DIAGCODE, second match is between unmatched codes after the first match, which then have extra zeros removed from the first or last two digit positions, with SASDATA.DIAGCODE.***

Obviously, there is a certain increase in the percentage of paid services associated with the 2nd match of the diagnostic codes in terms of total services (1.4 percent increase overall). The significant increases for five digit numeric codes (69 percent increase) and X-code (13.6 percent increase) are worth noting.

4.3 Unmatched Diagnostic Codes and Practitioners

Section 4.2 revealed that certain diagnostic codes in the PMEMASTER are invalid because they somehow have extra zeros in the codes. They might be valid if these zeros were removed. The possible explanation of this is that either physicians miscode diagnoses, or the Teleplan systems, (used by physicians to electronically submit their claims) somehow automatically add zeros in the front or at the back of the codes. To investigate this, further analyses were performed to see if there is a group of physicians who consistently submit these types of unmatched codes. If a group is identified, the Teleplan systems used by these physicians will be checked out, through the MSP Claims System, to determine if these systems are in fact adding the zeros.

There are a few of physicians who submit a high percentage of diagnostic codes of the type 00XXX (ie. five digit numeric codes with the first two digits as zeros), and the type XXX00 (five digit numeric codes with last two digits as zeros). Tables 4.5 (List of Physicians Who Submitted High Percentage of Diagnostic Codes as Type of 00XXX) and 4.6 (List of Physicians Who Submitted High Percentage of Unmatched Diagnostic Codes As Type of XXX00) lists a sample of these. Individual practitioners listed in Tables 4.5 and 4.6 submitted a significantly high proportion of these particular code types. For example, physician two in table 4.5 used a total of 402 diagnostic codes in fiscal year 1994/95, of which 400 (99.5 percent) are type 00XXX codes. It seems that there are identifiable groups of physicians who consistently submit some of these particular codes types.

Table 4.5: List of Physicians Who Submitted High Percentage of Diagnostic Codes As Type of 00XXX (Five Digit Numeric Codes with Two Zeros at Beginning), 1994/95

List of Individual Physician	Total Codes Submitted	Code 00XXX Counts	Percent	Total Paid Services	Paid Services Associated to 00XXX Code	Percent
1	19	19	100.0	4,669	4,669	100.0
2	402	400	99.5	9,722	9,719	100.0
3	55	54	98.2	3,276	3,275	100.0
4	26	25	96.2	6,944	6,942	100.0
5	17	16	94.1	9,540	9,537	100.0
6	23	21	91.3	3,865	3,863	99.9
7	63	54	85.7	5,211	5,179	99.4
8	96	81	84.4	11,247	9,868	87.7
9	48	37	77.1	3,314	3,249	98.0
10	221	117	52.9	8,345	7,673	91.9
11	248	99	39.9	8,297	4,495	54.2
12	37	14	37.8	1,880	700	37.2
13	217	73	33.6	2,480	299	12.1

4.4 Physicians' Teleplan Systems and Diagnostic Codes

Table 4.6: List of Physicians Who Submitted High Percentage of Diagnostic Codes As Type of XXX00 (Five Digit Numeric Codes Ended with Two Zeros), 1994/95

List of Individual Physician	Total Codes Submitted	Code XXX00 Counts	Percent	Total Paid Services	Paid Services Associated to XXX00 Code	Percent
1	287	211	73.5	7,315	5,037	68.9
2	369	269	72.9	9,367	7,020	74.9
3	232	167	72.0	10,934	5,258	48.1
4	193	138	71.5	6,682	4,005	59.9
5	251	179	71.3	14,583	8,999	61.7
6	292	204	69.9	8,771	5,419	61.8
7	216	149	69.0	3,880	2,465	63.5
8	73	50	68.5	5,805	3,987	68.7
9	72	49	68.1	11,486	7,566	65.9
10	61	41	67.2	264	176	66.7
11	247	164	66.4	15,259	10,457	68.5
12	247	152	61.5	5,392	3,469	64.3
13	120	73	60.8	6,699	4,739	70.7
14	163	99	60.7	18,484	6,910	37.4
15	103	61	59.2	6,436	5,298	82.3
16	168	90	53.6	704	437	62.1
17	257	132	51.4	5,697	2,254	39.6

A list of practitioners (hereafter refered to as designated practitioner), who submitted a high proportion of code types XXX00 and 00XXX, was provided to the MSP Clams Systems Branch in order for them to investigate if the Teleplan systems used by these physicians automatically attached zeros to the submitted diagnostic codes. As a result, four different types of Teleplan software were found to be suspect, and there are a total of 684 payees (containing 1,416 practitioners) who have been using these software.

In order to further clarify this suspicion, a random sampling from the remainder of the medical practitioners (total 6,146, excluding the above 1,416 practitioners) of an equivalent size was created as a control group for a comparison analysis. Finally, 1,463 practitioners were included in the control group.

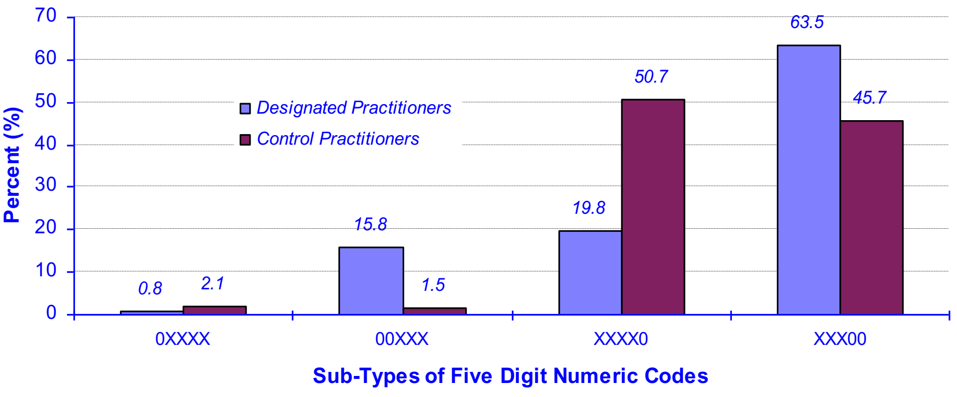
The analysis was performed in order to compare the percentage distribution of code frequency across the different code types between the designated and the control practitioner groups. Table 4.7 (Percent Distribution of Code Frequency by Types of Codes For Designated and Control Practitioners) shows the result which indicates that the percentage of code types 00XXX and XXX00 are higher for disignated practitioners (0.5 percent and 1.9 percent respectively) than the control ones (0.0 percent and 0.5 percent respectively). The difference does not seem to be significant. However, if the comparison focuses only on the five digit numeric codes, the difference is more obvious. Figure 4 (Percent Distribution of Frequency of Five Digit Numeric Codes by Sub-Types Between Suspect and Control Practitioners) is the comparison of percent distribution of code frequency by sub-types of five

Table 4.7: Percent Distribution of Code Frequency by Types of Codes For Designated and Control Practitioners*, 1994/95

Types of	Designated Practitioners		Control Practitioners	
Code	Code Frequency	Percent	Code Frequency	Percent
XXX__	147,438	69.6	160,578	80.8
XXXX_	15,657	7.4	24,099	12.1
XXXXX	810	0.4	1,192	0.6
0XXX_	843	0.4	1,221	0.6
XXX0_	40,754	19.2	9,553	4.8
0XXXX	53	0.0	42	0.0
00XXX	1,002	0.5	30	0.0
XXXX0	1,256	0.6	1,012	0.5
XXX00	4,020	1.9	914	0.5
Other Types	90	0.0	153	0.1
All Types	***211,923***	***100.0***	***198,794***	***100.0***

** Designated practitioners, provided by the Claims System Branch of MSP, have been using four types of Teleplan software which are suspected to add extra zeros to submitted diagnostic codes; the control practitioners were randomly selected from the remained practitioners.*

Figure 4: Percent Distribution of Frequency of Five Digit Numeric Codes by Sub-Types Between Designated and Control Practitioners, 1994/95



digit numeric codes. It is clear that the designated practitioners have a significantly higher percentage for code types 00XXX (15.8 percent) and XXX00 (63.5 percent) than the control practitioners (1.5 percent and 45.7 percent respectively).

In summary, it appears that some Teleplan software seem to have something of a problem with the digit zero. However, the evidence is not clear enough to support this because the four Teleplan software that are suspect also submit a high proportion of three digit numeric codes. Logically, if the software adds zeros to those diagnostic codes because they are not filling the full length of the variable DIAGCD (five digit character variable), then there should not be any codes left in forms of XXX__ and XXXX_, however this is not true. Therefore, there is probably something else, alone, or in conjunction with the Teleplan software, which creates these code types as recorded in the PMEMASTER.

5. Invalid Diagnostic Codes in the PMEMASTER - Estimate their impact

The impact of invalid diagnostic codes in the PMEMASTER on the accuracy of the diagnosis related information obviously depends on how the data is extracted and what code types are used to extract the information. According to the previous discussions and results already outlined, a rough estimate of the impact of invalid diagnosis codes in the PMEMASTER is summarized below.

5.1 Three Digit Numeric Codes

The impact is not significant because a full list of official three digits codes is covered in the PMEMASTER. Patient counts, paid services, and paid amounts associated with the matched three digits codes are approximately 100 percent of the totals associated with all three digits codes in the PMEMASTER (see table 2.3). In other words, invalid three digit numeric codes have little impact on information related to this type of code.

5.2 Four Digit Numeric Codes

The impact is minor because about 87 percent of standard four digit numeric codes are covered in the PMEMASTER. Patient counts, paid services, and paid amounts associated with the matched four digit numeric codes are over 85 percent of the totals associated with all four digits codes in the PMEMASTER (see table 2.3).

5.3 Five Digit Numeric Codes

The impact is significant because 70 percent of standard five digit numeric codes are covered in the PMEMASTER. Patient counts, paid services, and paid amounts associated with the matched five digit numeric codes are less than 10 percent of the totals associated with all five digits codes in the PMEMASTER (see table 2.3).

5.4 All Character Codes

The impact is minor because almost 100 percent of standard character codes are covered in the PMEMASTER except V-codes which are 67 percent covered. However, the patient counts, paid services, and paid amounts associated with the matched character codes are over 95% of the totals associated with all character codes in the PMEMASTER (see Table 2.3).

IV. Conclusions

The above analyses have demonstrated that the quality of diagnostic codes in the PMEMASTER is generally acceptable, especially that of three digit numeric codes and character codes. However, the methods used to extract different types of the codes (such as three, four, and five digits, etc.) will vary and have some impact on the accuracy of the related information. According to the analyses, the following methods are recommended to provide any information related to diagnostic codes from the PMEMASTER.

1) Summarize information, such as paid services and paid amounts, by individual diagnostic codes. The resultant dataset will be individual ICD code specific information.
2) If the information requested is related to the codes which have four or five digits subdivisions, say code 360 , *Disorders of the Globe*, the codes 360, 3600 to 3609, and 36000 to 36004, 36011 to 36019, 36020 to 36024 & 36029, 36030 to 36034, 36040 to 36044, 36050 to 36055 & 36059, 36060 to 36065 & 36069, 36081 & 36089 must be included because those four and five digits codes are subdivisions of three digit code 360. The extractions for this list of codes can then be done from the resultant data set.
3) If the information requested is related to the codes which only have four digit subdivisions, say 153, *Malignant Neoplasm of Colon*, the codes 153 and 1530 to 1539 must be included, and the extractions based on these codes can then be done from the resulted data set.

4) If the requested information is related to the character codes, such as L, X, Z, and V, the extraction is different. Firstly, create two sub-datasets by matching the resultant data set to SASDATA.DIAGCODE by DIAGCD; one sub-set will contain all matched data, and the other all non-matched. Using standard character codes such as 01L, 01X, and 01Z etc to directly extract information from matched sub-sets by "*WHERE*" clause, say *WHERE DIAGCD IN ('01L' 01Z' '01X')*; using "*SUBS.*" clause to extract information from unmatched sub-set, like *"WHERE SUBST.(DIAGCD, 1and ,3) in ('01L', '01X', '01Z')"*. The extracted information will be reasonably complete.
5) Five digit numeric codes are not generally recommended to provide information because of the incompleteness of the codes.

Although the quality of ICD codes in the PMEMASTER is generally acceptable, one serious problem is that most of the codes and associated paid services and paid amounts are reported by three digits codes which are major diagnostic group codes, and lack specificity. Therefore, the diagnostic information may not be appropriate for those applications which require more specific diagnostic information.