

Department of Health and Human Services

**OFFICE OF
INSPECTOR GENERAL**

**USING SOFTWARE TO DETECT
UPCODING OF HOSPITAL BILLS**



JUNE GIBBS BROWN
Inspector General

August 1998
OEI-01-97-00010

OFFICE OF INSPECTOR GENERAL

The mission of the Office of Inspector General (OIG), as mandated by Public Law 95-452, is to protect the integrity of the Department of Health and Human Services programs as well as the health and welfare of beneficiaries served by them. This statutory mission is carried out through a nationwide program of audits, investigations, inspections, sanctions, and fraud alerts. The Inspector General informs the Secretary of program and management problems and recommends legislative, regulatory, and operational approaches to correct them.

Office of Evaluation and Inspections

The Office of Evaluation and Inspections (OEI) is one of several components of the Office of Inspector General. It conducts short-term management and program evaluations (called inspections) that focus on issues of concern to the Department, the Congress, and the public. The inspection reports provide findings and recommendations on the efficiency, vulnerability, and effectiveness of departmental programs.

OEI's Boston office prepared this report under the direction of Mark R. Yessian, Ph.D., Regional Inspector General. Principal OEI staff included:

BOSTON REGION

Russell Hereford, *Project Leader*
Kenneth Price
Nicola Pinson

HEADQUARTERS

Mark Krushat
Tricia Davis
Linda Moscoe
Brian Ritchie

To obtain a copy of this report, please contact the Boston Regional Office by telephone at (617) 565-1050 or by fax at (617) 565-3751.

Reports are also available on the World Wide Web at our home page address:

<http://www.dhhs.gov/progorg/oei>

EXECUTIVE SUMMARY

PURPOSE

The purpose of this study is to test the ability of commercial software products to identify Diagnosis Related Group upcoding in Medicare hospital bills.

BACKGROUND

Since 1983, Medicare has paid acute care hospitals for the care of its beneficiaries under a prospective payment system using Diagnosis Related Groups (DRGs). In fiscal year 1996, expenditures for inpatient hospital care under this system totaled \$77.6 billion.

Improper Payments

Improper hospital payments are a continuing concern in the Medicare program. In its Chief Financial Officers audit of Medicare, the Office of Inspector General (OIG) estimates that in fiscal year 1997, \$4.1 billion of DRG payments were inappropriate due to lack of medical necessity, insufficient or no documentation, or incorrect coding.

One particular concern is upcoding of hospital bills, the practice of billing for a hospital stay more expensive than the one actually incurred. In previous studies, we found upcoding in DRGs ranging from 7 to 13 percent.

Commercial Upcoding Detection Software

Dozens of vendors now offer upcoding detection software that locates potentially upcoded DRGs by analyzing electronic files of hospital bills. These products are likely to become increasingly sophisticated as the state of the art of computing races ahead.

In this inquiry we evaluated the ability of two promising products to identify DRG upcoding. First, we used these products to identify hospital bills with suspected upcoded DRGs. Then we used professional record reviewers to perform a blinded medical review on a sample of cases to assess how well the products predicted DRG upcoding at the hospital, DRG, and case levels.

FINDINGS

Hospital Level

Hospitals identified by the software had an average upcoding rate of 11.5 percent, more than double the 5.3 percent average upcoding rate of the control hospitals.

However, the software also identified as high upcoders a substantial number of hospitals in which our medical record review identified few or no upcoded cases.

DRG Level

The software performed best at identifying upcoded cases in three DRGs that show the highest rates of actual upcoding: DRG 87, pulmonary edema and respiratory failure; DRG 79, respiratory infections and inflammations; and DRG 144, other circulatory system diagnoses. These three DRGs comprise 3.5 percent of all Medicare discharges, or about 350,000 discharges per year.

However, among the most commonly occurring DRGs, we found that the software was no more effective in identifying upcoded cases than among other DRGs.

Case Level

The software successfully identified between 50 and 60 percent of cases that were actually upcoded. Over 40 percent of upcoded cases went undetected.

However, only 10 to 20 percent of cases that the software identified as upcoded were, in fact, upcoded.

CONCLUSIONS

Our analysis of the software products provides some basis for optimism about the role that such products can play in detecting DRG upcoding. Yet we temper that optimism with strong caution as to the current state of the art of this software and the need to couple its use with other measures in the detection and prevention portfolio.

The software we examined showed modest success in identifying hospitals with a high rate of upcoding and upcoded cases within a narrowly defined group of DRGs that exhibited the most frequent upcoding. Thus, software could be used to identify hospitals that may need close scrutiny either before or after Medicare pays them. However, because these products were distinctly less successful for most other DRGs, we see only a limited role for these products at the current time.

It is likely, however, that the software market will continue to develop over time, and that products such as these will advance in sophistication and become more useful as part of a fraud detection strategy. No doubt HCFA will want to stay abreast of opportunities that this technology may present.

VENDOR COMMENTS

We provided copies of our draft report and our contractor's report to the three vendors whose software products we tested. We wish to express our appreciation to these

companies for their willingness to let us use their products in this inspection, and for their comments and analysis of our draft report.

These companies raised two general points in their responses. First, the companies indicated that their products could be modified in ways that address the Medicare population more directly, and that they are continuously updating and enhancing their products. We note that our purpose was not to develop new software, but to test commercially available off-the-shelf software. We did not modify the vendors' software, nor did we ask them to modify the software or to develop a specific software product for this purpose.

Second, they questioned the methods we used to test the software. We stand by our methodology. We tested the software in a way that we considered would be useful to an agency such as HCFA. We took the software's underlying individual claims based approach and aggregated the results of individual claims analysis to the provider level. We then verified the software products' performance by reviewing cases among a sample of the providers that the software identified as having a high rate of upcoding. In our judgement, this was a practical extension of the software. We used these products in a manner that might identify and focus on providers that bear additional scrutiny in a fraud prevention and detection effort.

We also address the methodological issues that one of the vendors raised in its response to the report.

We include the full text of each vendor's comments in Appendix D.

TABLE OF CONTENTS

	PAGE
EXECUTIVE SUMMARY	i
INTRODUCTION	1
FINDINGS	
Hospital Level	4
DRG Level	5
Case Level	7
CONCLUSIONS	8
VENDOR COMMENTS	9
APPENDICES	
A: Software Vendor Search	A-1
B: Testing Methodology	B-1
C: Statistical Tables	C-1
D: Text of Software Vendors' Comments	D-1
E: Endnotes	E-1

INTRODUCTION

PURPOSE

The purpose of this study is to test the ability of commercial software products to identify Diagnosis Related Group upcoding in Medicare hospital bills.

BACKGROUND

Since 1983, Medicare has paid hospitals for the care of its beneficiaries under a prospective payment system (PPS) using Diagnosis Related Groups (DRGs). In fiscal year 1996, expenditures for hospital care under this system totaled \$77.6 billion.¹ Under PPS, payment to hospitals for each Medicare case is based on a hospital-specific payment rate, multiplied by the weight of the DRG to which the case is assigned. Each DRG weight represents the average resources required to care for cases in that particular DRG relative to the average resources used to treat cases in all DRGs.

Cases are classified into DRGs based on the principal diagnosis, up to eight additional diagnoses, and up to six procedures performed during the stay, as well as the age, sex, and discharge status of the patient. Upon discharge, the physician summarizes information on a discharge face sheet. A hospital coder then reviews the entire medical record and uses that information to assign the most appropriate codes from the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). The hospital uses this information to prepare a claim for payment, which it forwards to the Medicare fiscal intermediary. The intermediary applies a series of edits to the claim, then groups the ICD-9-CM codes in the claim into the appropriate DRG for payment to the hospital.

Improper Payments

Improper hospital payments are a continuing concern for Medicare's Part A trust fund. In its Chief Financial Officers audit of Medicare, the Office of Inspector General (OIG) estimates that in Fiscal year 1997, \$4.1 billion of hospital payments were inappropriate due to lack of medical necessity, insufficient or no documentation, or incorrect coding.² One particular concern is upcoding of hospital bills, the practice of billing for a hospital stay more expensive than the one actually incurred. In previous OIG studies, we found upcoding in DRGs ranging from 7 to 13 percent.^{3,4}

Commercial Upcoding Detection Software

As the pressure on public and private insurers to eliminate improper payments has risen, the market for software to detect upcoding has experienced rapid growth. Dozens of vendors now offer such software. These products analyze the diagnostic and administrative data from each hospital bill in an electronic claims file to predict whether

the DRG contained in the bill is upcoded. Many vendors sell their products “off the shelf”— ready to be installed and utilized with minimal investment and setup time. These products are likely to become increasingly sophisticated as the state of the art of computing races ahead.

This Inquiry

Within its fiscal year 1996 Chief Financial Officers audit of Medicare, OIG recommended that HCFA enhance prepayment and postpayment controls by updating computer systems to better detect improper claims. In this inquiry we evaluated the ability of two promising products to identify DRG upcoding through electronic analysis of hospital bills. We chose these two products from a field of 21 vendors who offer similar products. We intended this test to be illustrative of how software might complement HCFA’s existing program integrity initiatives by functioning as one part of a broad strategy for DRG payment safeguarding. We based our evaluation on a blinded medical record review of a national sample of 2,622 Medicare cases from 1996. The review was performed by an independent contractor using accredited medical records professionals.

METHODOLOGY

We executed the test in two phases, using a contractor with expertise in statistical sampling and medical record review for highly specialized tasks.

In phase one we used our contractor to search for vendors of upcoding detection software. The search initially identified 57 vendors whose product description indicated some type of claim auditing software or services. Further research of these 57 vendors reduced the list to 21 vendors who had software that appeared relevant to our study. Out of these 21 software vendors, 3 agreed to participate in our study.⁵

Using software from these three vendors, we analyzed 100 percent of Medicare inpatient claims from January through June of 1996 to identify claims that appeared to be upcoded.⁶ Next, we collapsed the output from each software product to generate three lists of hospitals with high predicted rates of upcoding. Through correlation analysis, we discovered a strong relationship between the lists of hospitals generated by two of the software products, while the list from the third product differed significantly. Due to limits on the number of medical records we could review for this study, we decided to focus our inquiry by testing only the output from the two software products whose lists were closely correlated.⁷ Therefore, as the first step of our sample, we selected 50 hospitals that both products predicted had high rates of upcoding. We refer to this group of hospitals as our test sample.

As a control, we also selected a sample of 20 hospitals that fell into similar size strata as our test sample but did not have high predicted rates of upcoding. This brought the total number of hospitals in our study to 70 — 50 hospitals with high predicted rates of upcoding and 20 hospitals without high predicted rates of upcoding. From each hospital,

we sampled 40 Medicare inpatient admissions billed under any of the 50 most prevalent DRGs in 1996. This brought the total number of cases in our study to 2,800. We were able to obtain and complete analysis on 2,622 (94 percent) of these cases. Tables C-1 through C-4 in Appendix C present data on the characteristics of the hospitals and cases examined in this review.

In phase two our contractor performed a blinded medical record review of each case. For this task, the contractor used Registered Records Administrators, Accredited Records Technicians, Certified Coding Specialists, and physicians. Based on the contents of the medical record, the contractor derived a new set of ICD-9-CM diagnostic and procedure codes and used them to generate a new DRG for each case. If there was discrepancy between the new DRG and the DRG for which the hospital had billed Medicare, the case was referred for a second blind review to determine the final DRG. If the contractor calculated a final DRG that was less expensive than the DRG the hospital billed Medicare for, we defined a case as upcoded.⁸ Thus, for each of the 2,622 cases in our review, we knew which had been properly coded and which had been upcoded.

We then used this information to determine if the two software products successfully predicted whether each case in our sample had an upcoded DRG. We analyzed these data on three levels: by hospital, by DRG, and by case. To perform hospital-level and DRG-level analysis, we aggregated our data by hospital and DRG to compare actual and predicted rates of upcoding. Case-level analysis examined the success of the software in predicting DRG upcoding on a case-by-case basis. Our analyses used t-tests and logistical regression to determine statistical differences.

A detailed description of the software vendor search and testing methodology appears in Appendices A and B.

We conducted this study in accordance with the *Quality Standards for Inspections* issued by the President's Council on Integrity and Efficiency.

FINDINGS

Hospital Level

Hospitals identified by the software had an average upcoding rate of 11.5 percent, more than double the 5.3 percent average upcoding rate of the control hospitals.

The 50 hospitals in the test group — those hospitals that the software identified as having high rates of upcoding — did in fact exhibit higher upcoding rates than the hospitals in our control sample. The 12 hospitals in which our medical records reviewers found the highest upcoding rates were in the test sample. Seven of these test hospitals had upcoding rates of 25 percent or higher; more than half (26 out of 50) had an upcoding rate of 10 percent or higher. (See Table I.)

The 20 hospitals in the control sample tended to have low upcoding rates. In the sample of cases from control hospitals, our reviewers found upcoding rates below 5 percent in half the hospitals. They found no upcoded cases at all in 5 of the 20 control hospitals.

We performed logistic regression analysis to control for the effects of additional variables related to the hospitals (*e.g.*, teaching status, ownership) and individual cases (*e.g.*, patient gender and age). Even taking these variables into account, we found that the likelihood of a case being upcoded in hospitals identified by the software was almost twice as high as it was for hospitals in the control sample. Table C-5 in Appendix C presents the full results of this analysis.

Table I: Upcoding Rates for Test and Control Hospitals			
	Entire Sample (n=70)	Test Hospitals (n=50)	Control Hospitals (n=20)
Average upcoding rate (<i>t</i> =3.57 <i>p</i> <.001)	9.8%	11.5%	5.3%
Number (percent) with upcoding ≥25%	7 (10%)	7 (14%)	0 (0%)
Number (percent) with upcoding ≥10% and <25%	22 (31%)	19 (38%)	3 (15%)
Number (percent) with upcoding <10%	41 (59%)	24 (48%)	17 (85%)

However, the software also identified as high upcoders a substantial number of hospitals in which our medical record review identified few or no upcoded cases.

In 6 of the 50 hospitals in the test sample, our medical records reviewers found no upcoded cases, even though the software predicted that these hospitals would have high

rates of upcoding. In 15 of these 50 hospitals we found upcoding rates of 5 percent or less.

At the same time, it is worth noting that our reviewers found upcoding rates of 10 percent or higher in 3 of the 20 control hospitals.

DRG Level

Measuring the Effectiveness of Software in Identifying Upcoded Cases

The effectiveness of a software product can be measured along two dimensions, referred to as sensitivity and specificity. Each dimension can be expressed as a percentage.

Sensitivity measures the extent to which the software identifies all cases that have been upcoded. In our sample of 2,622 cases, our independent medical records reviewers determined that 254 cases (9.7 percent) had been upcoded. A software product that was perfectly sensitive would identify all 254 of these cases.

Specificity assesses the software's efficiency. Specificity measures the software's ability to discriminate between those cases that were upcoded and those cases that were not upcoded, *i.e.*, the extent to which the software identifies only those cases that really were upcoded. If the software were perfectly specific, every case that it identified would be upcoded.

An ideal product would be perfectly sensitive and perfectly specific — in our review, for example, a perfect product would have selected all 254 upcoded cases and omitted the other 2,368 cases.

In reality, there often is a trade-off between sensitivity and specificity: To achieve greater sensitivity, the software must cast a wide net; this means that it might identify some cases that were not really upcoded, referred to as “false positives.” Conversely, to achieve greater specificity, the software risks missing some cases that actually were upcoded; those cases that it misses are referred to as “false negatives.”

We examined the performance of the software among two sets of DRGs which we consider potentially high risk to the Medicare program in terms of potential dollars lost: those DRGs in which we found a high level of upcoding and those DRGs which occur most frequently. We examined these DRGs to determine whether the software might be most efficiently utilized by identifying a subset of DRGs that represent potentially high cost to the Medicare program, either because they exhibited high rates of upcoding or because of the sheer volume of cases.

The software performed best at identifying upcoded cases in DRGs that show the highest rates of actual upcoding.

The software was most accurate in identifying cases in the three DRGs with the highest actual rates of upcoding. These 3 DRGs comprise 3.5 percent of all Medicare discharges, or about 350,000 discharges per year:

DRG 87 (Pulmonary edema and respiratory failure). Our medical records reviewers found an actual upcoding rate of 41 percent. The software products had sensitivity rates of 69 percent and 61 percent, and specificity rates of 60 percent and 47 percent.

DRG 79 (Respiratory infections and inflammations, age > 17 with complications or comorbidities). Our medical records reviewers found an actual upcoding rate of 35 percent. The software products had sensitivity rates of 95 percent and 55 percent, and specificity rates of 36 percent each.

DRG 144 (Other circulatory system diagnoses with complications or comorbidities). Our medical records reviewers found an actual upcoding rate of 30 percent. The software products had sensitivity rates of 86 percent and 71 percent, and specificity rates of 86 percent and 42 percent.

For DRGs with lesser—but still high—rates of upcoding, however, the software was less accurate. For example, one software product flagged no cases in the DRGs with the fourth highest upcoding rate (DRG 239, with 24 percent actual upcoding) or the fifth highest upcoding rate (DRG 429, with 23 percent upcoding); the other product was only slightly more successful. For informational purposes, we present data on sensitivity and specificity of the software for the 10 DRGs with highest rates of upcoding in Appendix C, Table C-6.

One implication arising from this analysis is that once DRGs that exhibit high levels of upcoding have been found — for example, through ongoing case review and analysis of discharges — the software products may have a role to play in helping to identify specific cases within those DRGs that merit further scrutiny.

We also examined how well the software performed in detecting case-specific upcoding among the group of 10 DRGs with the highest upcoding rates, versus the other DRGs we reviewed in this inspection. (See Table C-7 in Appendix C.) We found no statistical difference between these two groups in the software’s sensitivity (*i.e.*, its ability to identify upcoded cases). We did, however, find that the software was more specific among those frequently upcoded DRGs. In other words, those cases that the software did identify tended to be actually upcoded.

However, among the most commonly occurring DRGs, we found that the software was no more effective in identifying upcoded cases than among other DRGs.

The 10 most commonly occurring DRGs comprise 10 percent of all Medicare discharges, or about 1 million discharges per year. Within our sample, they comprised 13 percent of

cases reviewed, yet they accounted for 53 percent of upcoded cases.⁹ Table C-8 in Appendix C provides data on these 10 most commonly occurring DRGs.

We examined how well the software performed in detecting case-specific upcoding among this group of 10 DRGs, versus the other DRGs we reviewed in this inspection. (See Table C-9 in Appendix C.) We found no statistical difference in sensitivity or specificity for either product in its ability to detect upcoding among these 10 most commonly occurring DRGs, compared with the performance of the products in correctly identifying upcoded cases in other DRGs.

Case Level

Our sensitivity analysis showed that the software products successfully identified between 50 and 60 percent of cases that were actually upcoded.

Our medical records reviewers determined that our sample contained 254 cases that had been upcoded. Of these 254 upcoded cases, one product identified 133 (52 percent) of these cases, and the other product identified 147 cases (58 percent). This sensitivity rate has an important implication: over 40 percent of cases that actually were upcoded went undetected by these products.

However, our specificity analysis showed that only 10 to 20 percent of cases that the software products identified as upcoded were, in fact, upcoded.

One product identified 685 cases as upcoded, but only 133 (19 percent) of these cases were determined by our reviewers to be upcoded. For the other product, out of 1,284 cases it identified as upcoded, 147 (11 percent) were determined by our reviewers to actually be upcoded. Such a low specificity rate reduces the efficiency of the software as a detection tool by requiring that multiple cases be reviewed in order to locate each upcoded case. In essence, for the product with 19 percent specificity, reviewers would need to examine 4 false leads to find 1 case that truly was upcoded. For the product with the 11 percent specificity rate, that review level rises to 9 false leads for each truly upcoded case.

CONCLUSIONS

The software we examined showed modest success in identifying hospitals with a high rate of upcoding. Those hospitals that the software identified were twice as likely to have upcoded cases as a control group of hospitals. This finding leads us to believe that the software could be used in an ongoing way to identify hospitals that are likely to upcode their Medicare bills.

There are various approaches as to how the software might be applied at the hospital level. For example, HCFA might wish to use this software as a tool in its post-payment recovery efforts. Based on the results we found, HCFA could use such software to retrospectively identify hospitals in which it would be likely to find a high level of upcoded cases and commensurate overpayment. Alternatively, the agency could use this software to identify hospitals that have previously demonstrated a tendency to upcode, and then perform focused review on cases from these hospitals prior to making payments.

The software also showed some success within the narrowly defined group of DRGs that exhibited the most frequent upcoding. Because the software was relatively successful in identifying particular cases that were upcoded among these DRGs, its use here could be expected to yield significant returns. For post-payment recovery efforts, HCFA could opt to focus on cases in the upcoded DRGs; analogously, the software could be used prospectively to identify cases in particular DRGs for review prior to payment.

At the same time, our review leads us to raise caution about these products, particularly at the individual case level. While they worked well for the most frequently upcoded DRGs, our review determined that these products were distinctly less successful for other DRGs. It is for this reason that we see only a limited role, as described above, for these products at the current time.

The two software products that we reviewed were illustrative of what was available on the market in the Spring of 1997. We believe, however, that it is likely that the software market will continue to develop, and that products such as these will advance in sophistication and expand in their usefulness as part of a fraud detection strategy. No doubt HCFA will want to stay abreast of opportunities that this technology may present.

VENDOR COMMENTS

We provided copies of our draft report and our contractor's (FMAS) report to the three vendors whose software products we tested. We wish to express our appreciation to these companies for their willingness to let us use their products in this inspection, and for their review of and comments on our draft report. Their overall comments reflect support for analytical work of this nature; however, the vendors express some concerns about our application of the software and raise some questions about our methodology. We address their comments here, and we include the complete text in Appendix D.

We wish to address two points that the vendors raised regarding the manner in which we applied the software. First, each vendor indicates that it is continuously updating and enhancing its products. One company even notes specifically that the software system we tested could be modified in ways that address the Medicare population more directly. We are confident that software enhancements undoubtedly will continue to expand the potential for products such as these to play an important role in fraud prevention and detection.

The purpose of this inspection, however, was not to develop new software, but to test commercially available off-the-shelf software. Our interest was in determining if products that were on the market at the time we conducted our review (Spring 1997) could prove useful in identifying hospitals that showed a high rate of DRG upcoding. Consequently, we did not modify the vendors' software, nor did we ask them to modify the software or to develop a specific software product for this purpose. Rather, we utilized the vendors' software packages on an "as-is" basis.

Second, the vendors raised concerns about our use of the software to go beyond identification of individual claims that may have been coded incorrectly. We recognize that, to some extent, our test was a modification of the original intent of these software products, which is to detect specific clinical claims that are questionable. In essence, we took this underlying approach and extended it. We aggregated the results of the individual claims analysis in order to identify hospitals that the software showed have a tendency toward upcoding. We then verified the software products' performance by reviewing cases among a sample of the providers that the software identified as having a high rate of upcoding. In our judgement, this was a logical extension of the software to the practical realities of how it could be used in the Medicare program. We used these products in a manner that might identify and focus on providers that may bear additional scrutiny in a fraud prevention and detection effort.

We also wish to address the methodological issues that Dhrystone Systems raised in its response to the report. First, this vendor questions the methods that we used to select our sample, in that the sample contains outliers. We stand by our methodology; indeed we designed the methodology specifically to concentrate on the outlying providers that are most problematic. The experimental group comprised hospitals that the software identified as lying at least two standard deviations above the mean in the proportion of upcoding identified by the software. The control group comprised all remaining hospitals.

Second, Dhrystone also questions the appropriateness of our limiting the universe of cases reviewed to the 50 most common DRGs. In response, we note that we selected records from these 50 DRGs to focus our review effort. These 50 DRGs comprise nearly 70 percent of all discharges, and over 60 percent of all Medicare PPS reimbursement. Consequently, targeting these 50 DRG strikes us as a prudent means of focusing on where the greatest concentration of Medicare dollars lies. We do not generalize these results to broader populations of DRGs or of hospitals.

Dhrystone also states that the study purports to have been conducted in a double blind manner, and questions whether we did, in fact, do so. In response, we note that our review was conducted in a blinded manner; but we do not claim it was a double blind study. The initial review was conducted by a registered records coder in a fully blinded manner. If the coder found a discrepancy, the record was then unblinded; the coder then compared the hospital's reasoning with her reasoning, and arrived at a determination of the appropriate coding. If disagreement persisted between the coder and the hospital, a second blind review was conducted, and the results of both reviews compared. In essence, this is a conservative way of conducting a review such as this. It clearly gives the benefit of any initial doubt to the hospital. We consider such a conservative approach to be prudent and likely reflective of any practical application of such software by HCFA and the Office of Inspector General.

APPENDIX A

Software Vendor Search

Our contractor, FMAS Corporation, consulted with the World Development Group (WDG) to search for commercial software vendors that had products designed to locate DRG upcoding using claims data.^{10,11}

To begin the search for software vendors, WDG identified and interviewed relevant experts and companies to obtain names of probable software vendors, additional experts, and any other relevant information to assist in the search. Initially, WDG identified 33 experts in health informatics, medical expert systems, electronic medical records, and Medicare Part A claims payment.

WDG sent each expert a fax describing the purpose of the project, the software of interest, and the questions that it would ask in a telephone interview. WDG contacted and interviewed 25 of the experts. This effort resulted in the identification of 11 additional experts, 5 of which WDG interviewed. In total, WDG interviewed 30 experts.

Concurrent with interviewing experts, WDG conducted a literature and Internet search to identify relevant software vendors. WDG searched the following print sources:

- 1996 Annual Market Directory Issue*. Health Management Technology. 1996.
- Medical Hardware and Software Buyer's Guide. M.D. Computing 1995; 12 (6).
- Ankrapp, Betty (ed). Health Care Software Sourcebook. Gaithersburg, MD: Aspen Publishers, Inc., 1996.
- Frisch, Bruce (ed). The HCP Directory of Medical Software. Brooklyn, NY: Healthcare Computing Publications Inc., 1996.

The literature and Internet searches and interviews with industry experts identified 57 vendors whose product description indicated some type of claim auditing software or services. WDG faxed a letter to each of these vendors to determine if they sold a product that met the project's criteria of relevance. This search identified three more vendors who claimed to have a relevant product.

In total, 21 of the 57 vendors appeared to meet the initial criteria of relevance. In preparation for the telephone interview, WDG sent a fax to these vendors.

WDG interviewed 20 of the 21 vendors. One vendor did not respond to repeated calls. Of the 20 vendors interviewed, 6 confirmed having a relevant product. Interviews with the 6 confirmed vendors lasted an average of an hour. During these interviews, WDG requested brochures and any other available product literature, as well as a contact name for the software testing phase of the study.

Of the six vendors, five agreed to a test of their software with certain conditions. WDG conducted follow-up interviews to obtain client references and discuss the test that would be conducted. In preparation of these follow-up interviews, WDG developed questions to query vendors about their software, clients, and willingness to test their software. WDG sent each of the 5 vendors a fax describing the purpose of the test and the topics to be discussed during the follow-up interview.

WDG contacted and interviewed all 5 vendors. Each interview lasted an average of 15 minutes. During these interviews, WDG requested as references the names of two payers or fiscal intermediaries. If the vendor did not have payer or fiscal intermediary references, WDG accepted any client references. Subsequently, WDG interviewed two client references for an average of 10 minutes each.

Vendors' concerns about the test fell into three categories: 1) the size of our test (5-10 million claims records) was too large; 2) vendors were uncertain about how OIG would utilize the results of the test; and 3) OIG's desired layout of the output was not clear enough.

Because of these concerns, only three vendors chose to remain in the study and participate in a test of their software.

APPENDIX B

Testing Methodology

We executed the study in two phases, utilizing a contractor with expertise in medical record review and statistical sampling for highly specialized tasks. In phase one, we located software products that might detect upcoding, used these products to generate a sample of hospitals, and drew a sample of medical records from these hospitals. In phase two, we performed a DRG validation on each case in our sample and used the results of this validation to determine if the software products used in stage one accurately predicted DRG upcoding.

We began phase one by issuing a Request for Proposals to locate a contractor with expertise in medical record review and statistical sampling to assist in the study. We contracted with FMAS Corporation, a company with extensive experience performing case review and analysis for the health care programs of the U.S. Department of Health and Human Services and the Department of Defense.¹²

FMAS worked with World Development Group (WDG) to locate vendors of software that detects DRG upcoding. From a field of 57 probable vendors, WDG identified 3 vendors that had relevant software and were willing to participate in our test. (See Appendix A).

Sample Selection

We used the software from these three vendors to process 100 percent of Medicare Prospective Payment System (PPS) cases from January through June 1996.¹³ As output, each software flagged cases that it deemed likely to have an upcoded DRG. Next, we made 3 lists of hospitals with high predicted rates of upcoding by collapsing each software's output by hospital. Through correlation analysis, we discovered a strong relationship between the lists of hospitals from two of the software, while the list from the third software differed significantly. This meant that we would have to draw two separate samples to have a sample of hospitals that was representative of hospitals identified by all three software. Thus, due to limits on the number of medical records we could review for this study, we decided to focus our inquiry by testing only the output from the two software whose lists were closely correlated.¹⁴

To build our experimental (test) sample, we first selected hospitals that either of the two software indicated had a predicted upcoding rate of the mean rate plus two standard deviations. This process led to a group of 299 hospitals, which we stratified into three groups according to number of Medicare discharges in the 6-month file we analyzed: 300 or fewer discharges, 301 to 1,000 discharges, and over 1,000 discharges. Next, in

proportion to the total number of hospitals in each stratum, we randomly selected a total of 50 hospitals from across the 3 strata.

From each hospital, we then randomly selected 40 Medicare cases billed under any of the 50 DRGs that were most commonly used across the country during fiscal year 1996. As a control sample, we executed the same sampling strategy to select 800 cases from 20 hospitals that had did not have high predicted rates of upcoding. This brought our total sample to 2,800 cases: 2,000 of which were from hospitals that had high predicted rates of upcoding, 800 of which were from hospitals that did not have high predicted rates of upcoding. We then merged claims data from each case against Medicare's Enrollment Data Base (EDB) to obtain beneficiary name and the Online Survey Certification and Reports (OSCAR) system to obtain hospital name and address. We used this information to mail medical record request letters and case listings to the administrator of each hospital in our sample. Hospitals sent medical records to the OIG, where we logged them, gave them a quality check, and assigned each a tracking number. We then sent the records to FMAS for DRG coding validation.

DRG Coding Validation

During phase two of the study the contractor, FMAS, performed a DRG coding validation on 2,622 (94 percent) of the 2,800 records in our sample. FMAS, using Registered Records Analysts and Accredited Records Technicians, performed a blinded record review, in which the original ICD-9-CM and DRG codes were hidden. This review generated new ICD-9-CM codes and a new DRG code for each case in the sample. When FMAS completed reviewing a record, it compared the new codes to the previously hidden codes used by the hospital. Below is the DRG reconciliation process:

If FMAS' codes and the hospital's codes matched, FMAS noted the DRG as correctly coded by the hospital. Depending on the specific ICD-9-CM codes assigned by FMAS, it assigned one of the following two reconciliation reason codes to the case:

1. Confirm: Face Sheet, UB-92, FMAS codes and DRGs match.
2. DRGs match, but there is some variance in codes.

If FMAS' codes initially disagreed with those of the hospital, FMAS still noted the hospital's DRG as correctly coded by the hospital if its reviewer agreed with the hospital's coding after performing an unblinded reconciliation review. FMAS' reviewer then assigned one of the following reconciliation codes to the case:

3. DRGs differed because more than one diagnosis could have been the principal diagnosis according to guidelines and hospital selected principal diagnosis leading to lower-weighted DRG. FMAS did not recode or regroup these cases either in software or on its hardcopy worksheet.

4. DRGs differed because of a judgement-call situation not covered by guidelines or Coding Clinic. FMAS' reviewer gave the hospital the benefit of the doubt. FMAS did not recode or regroup these cases either in software or on its hardcopy worksheet.
5. DRGs differed but FMAS' reviewer, upon reviewing hospital's codes/DRG, noted that the hospital's DRG was correct. This was the only reconciliation reason category that FMAS recoded and regrouped in software and on the hardcopy worksheet so that its final DRG matched the initial hospital DRG.
6. UB-92 DRG differed but hospital face sheet matched FMAS' DRG. FMAS did not recode or regroup these cases either in software or on its hardcopy worksheet. This category was selected whenever the codes on the face sheet would have led to the same DRG as the FMAS DRG, but the UB-92 DRG and related codes were different.
7. FMAS reserved this reconciliation code for potential additional reconciliation reasons, but did not use it during the study.

Whenever the DRGs differed after reconciliation, FMAS assigned the following reconciliation reason code to the case:

8. DRGs differ. Upon review of hospital's DRG codes, FMAS' reviewer confirmed that FMAS' DRG was correct based upon coding guidelines and *Coding Clinic*. FMAS did not recode or regroup these cases either in software or on its hardcopy worksheet. FMAS recorded all applicable DRG variance reasons and one DRG variance type (described below) on its DRG variance worksheet. FMAS then completed a second blinded review of the case using a different reviewer.

Variance types for reconciliation reason 8:

Misspecification: The narrative principal diagnosis, a secondary diagnosis, or a procedure is not supported by the medical record.

Miscoding: The medical records department selected an incorrect ICD-9-CM numeric code for a correct narrative diagnosis or procedure.

Resequencing: The hospital substituted a secondary diagnosis for the correctly attested and coded principal diagnosis.

Other: The hospital made another type of error (such as incorrect discharge status) that led to DRG variance but cannot be categorized as numbers 1-3 above.

OIG Analysis

FMAS sent data for the completed medical record reviews to OIG in electronic format, keyed by our tracking number. We merged these data with the original inpatient claims data and additional administrative data to create our analytical files for the study.

We analyzed these files on three levels: by hospital, by DRG, and by case. To perform hospital-level and DRG-level analyses, we aggregated our data by hospital and DRG to compare actual and predicted rates of upcoding. Case-level analysis examined the success of the software in predicting DRG upcoding on a case-by-case basis. We used t-tests and logistical regression to determine statistical differences. We performed data analysis using SAS software.¹⁵

APPENDIX C

Statistical Tables

TABLE C-1					
CHARACTERISTICS OF HOSPITALS REVIEWED					
	Control Sample (n=20)		Test Sample (n=50)		Total Sample (n=70)
	n	(%)	n	(%)	n(%)
Number of Beds					
1-99	14	(70.0)	42	(84.0)	56 (80.0)
100-299	5	(25.0)	4	(8.0)	9 (12.9)
300+	1	(5.0)	4	(8.0)	5 (7.1)
Teaching Status					
Teaching	3	(15.0)	10	(20.0)	13 (18.6)
Nonteaching	17	(85.0)	40	(80.0)	57 (81.4)
Location					
Metropolitan	11	(55.0)	10	(20.0)	21 (30.0)
Nonmetropolitan	9	(45.0)	40	(80.0)	49 (70.0)
Control					
For profit	3	(15.0)	3	(6.0)	6 (8.6)
Nonprofit	14	(70.0)	20	(40.0)	34 (48.6)
Government	3	(15.0)	27	(54.0)	30 (42.9)
Number of Discharges, 1/96-6/96					
1-300	9	(45.0)	32	(64.0)	41 (58.6)
301-1000	8	(40.0)	13	(26.0)	21 (30.0)
1001+	3	(15.0)	5	(10.0)	8 (11.4)
Source: OIG analysis of the FY 1996 Medicare Provider Analysis and Review (MEDPAR) file and data from the Online Survey Certification Reports (OSCAR) system.					

TABLE C-2 COMPARISON OF TEST SAMPLE WITH ALL HOSPITALS WITH HIGH PREDICTED RATES OF UPCODING				
	Test Sample (n=50)		Total High Predicted Group (n=299)	
	n	(%)	n	(%)
Number of Beds				
1-99	42	(84.0)	219	(73.2)
100-299	4	(8.0)	49	(16.4)
300+	4	(8.0)	31	(10.4)
Teaching Status				
Teaching	10	(20.0)	62	(20.7)
Nonteaching	40	(80.0)	237	(79.3)
Location				
Metropolitan	10	(20.0)	89	(29.8)
Nonmetropolitan	40	(80.0)	210	(70.2)
Control				
For profit	3	(6.0)	30	(10.0)
Nonprofit	20	(40.0)	124	(41.5)
Government	27	(54.0)	145	(48.5)
Number of Discharges, 1/96-6/96				
1-300	32	(64.0)	155	(51.8)
301-1,000	13	(26.0)	109	(36.5)
1,001+	5	(10.0)	35	(11.7)
Source: OIG analysis of the FY 1996 Medicare Provider Analysis and Review (MEDPAR) file and data from the Online Survey Certification Reports (OSCAR) system.				

**TABLE C-3
HOSPITAL CHARACTERISTICS BY CASE CHARACTERISTICS
FOR CASES REVIEWED**

	Control Sample (n=744)		Test Sample (n=1,878)		Total (n=2,622)	
	n	(%)	n	(%)	n	(%)
Number of Beds						
1-99	531	(71.4)	1,583	(84.3)	2,114	(80.6)
100-299	175	(23.5)	156	(8.3)	331	(12.6)
300 +	38	(5.1)	139	(7.4)	177	(6.8)
Teaching Status						
Teaching	108	(14.5)	372	(19.8)	480	(18.3)
Nonteaching	636	(85.5)	1,506	(80.2)	2,142	(81.7)
Location						
Metropolitan	407	(54.7)	383	(20.4)	790	(30.1)
Nonmetropolitan	337	(45.3)	1,495	(79.6)	1,832	(69.9)
Control						
For profit	119	(16.0)	120	(6.4)	239	(9.1)
Nonprofit	511	(68.7)	770	(41.0)	1,281	(48.9)
Government	114	(15.3)	988	(52.6)	1,102	(42.0)
Source: OIG analysis of the FY 1996 Medicare Provider Analysis and Review (MEDPAR) file and data from the Online Survey Certification Reports (OSCAR) system.						

TABLE C-4						
BENEFICIARY CHARACTERISTICS FOR CASES REVIEWED						
	Control Sample (n=744)		Test Sample (n=1,878)		Total (n=2,622)	
	n	(%)	n	(%)	n	(%)
Age (years)						
<65	106	(14.3)	268	(14.3)	374	(14.3)
65-74	196	(26.3)	489	(26.0)	685	(26.1)
75-84	269	(36.2)	695	(37.0)	964	(36.8)
85+	173	(23.3)	426	(22.7)	599	(22.9)
Sex						
Male	313	(42.1)	827	(44.0)	1,140	(43.5)
Female	431	(57.9)	1,051	(56.0)	1,482	(56.5)
Race						
White	648	(87.1)	1,573	(83.8)	2,221	(84.7)
Black	59	(7.9)	201	(10.7)	260	(9.9)
Other	28	(3.8)	87	(4.6)	115	(4.4)
Unknown	9	(1.2)	17	(0.9)	26	(1.0)
Source: OIG analysis of the FY 1996 Medicare Provider Analysis and Review (MEDPAR) file.						

**TABLES C-5
RESULTS OF LOGISTIC REGRESSION MODEL**

TABLE C-5A ODDS RATIOS ESTIMATES FOR STATISTICALLY SIGNIFICANT VARIABLES			
Variable	Estimate	90% Confidence Interval	
		Lower	Upper
Facility in Selected Group	1.94	1.41	2.66
Teaching Hospital	0.52	0.37	0.74
Publicly Owned	1.79	1.39	2.32
High Case Mix Index	2.59	1.96	3.43
Male	0.68	0.54	0.87
Age 75 to 84	1.40	1.07	1.85
Age 85+	1.60	1.18	2.17
Flagged by Product A	3.49	2.73	4.46
Flagged by Product B*	1.25	0.99	1.58

*This variable was not significant.

TABLE C-5B DEPENDENT VARIABLE IN LOGISTIC REGRESSION MODEL	
Case Actually Upcoded as Determined by Our Medical Records Reviewers	
Change	1 = DRG Upcoded (N=254) 0 = DRG not Upcoded (N=2,368)

**TABLE C- 5C
INDEPENDENT VARIABLES IN LOGISTIC REGRESSION MODEL**

	Facility Characteristics	
Profit	1 = For profit	0 = Non profit
Public	1 = Public	0 = Non profit
Location	1 = Nonmetropolitan	0 = Metropolitan
Teaching	1 = Teaching	0 = Nonteaching
Smallbed	1 = 1-99 beds	0 = 100-299 beds
Bigbed	1 = 300 + beds	0 = 100-299 beds
FewDC	1= 300 or fewer discharges	0 = 301 - 1,000 discharges
ManyDC	1 = Over 1,000 discharges	0 = 301 - 1,000 discharges
LowCMI	1 = CMI less than 0.9	0 = CMI between .9 and 1.1
High CMI	1 = CMI over 1.1	0 = CMI between .9 and 1.1
ExpSamp	1= Facility in test group of hospitals 0=Facility in control group of hospitals	
	Patient Characteristics	
Gender	1 = Male	0 = Female
Black	1 = Black	0 = White
Other	1 = Other	0 = White
Unknown	1 = Unknown	0 = White
Young	1 = Under 65	0=65-74
Seven5	1 = 75-84	0=65-74
Eight5	1 = 85 and Older	0=65-74
	Case Characteristics	
Surgical	1 = Surgical Claim	0 = Nonsurgical Claims
	Software Characteristics	
A_Hit	1 = Flagged by Software A 0=Not flagged by Software A	
B_Hit	1= Flagged by Software B 0= Not flagged by Software B	

TABLE C-6 SOFTWARE PERFORMANCE ON THE 10 DRGs WITH HIGHEST RATES OF UPCODING						
DRG (% of Medicare discharges)	Cases Reviewed (% of reviewed cases)	Percent Upcoded	Product A		Product B	
			Sensitivity	Specificity	Sensitivity	Specificity
87 Pulmonary edema & respiratory failure (0.6%)	32 (1.2%)	41%	69%	60%	62%	47%
79 Respiratory infections & inflammations age >17 w/cc (2.2%)	186 (7.1%)	35%	95%	36%	55%	36%
144 Other circulatory system diagnoses w/cc (0.7%)	23 (0.9%)	30%	86%	86%	71%	42%
239 Pathological fractures & musculoskeletal & conn tiss malignancy (0.5%)	25 (1.0%)	24%	0%	0%	33%	18%
429 Organic disturbances & mental retardation (0.4%)	13 (0.5%)	23%	0%	N/A*	33%	50%
416 Septicemia age >17 (2.0%)	84 (3.2%)	20%	94%	23%	82%	22%
475 Respiratory system diagnosis with ventilator support (0.9%)	26 (1.0%)	19%	100%	20%	100%	26%
188 Other digestive system diagnoses age >17 w/cc (0.6%)	17 (0.6%)	18%	0%	0%	67%	20%
121 Circulatory disorders w/AMI & C.V. comp disch alive (1.5%)	54 (2.1%)	15%	25%	15%	88%	15%
316 Renal failure (0.8%)	34 (1.3%)	15%	20%	33%	20%	6%

*Note:

Sensitivity = N/A when we found no upcoded cases within a DRG, *i.e.*, the denominator in our sensitivity calculation is zero.

Specificity = N/A when the software did not flag any cases within a DRG, *i.e.*, the denominator of our specificity calculation is zero.

**TABLE C-7
SOFTWARE PERFORMANCE ON TEN DRGs WITH HIGHEST RATES OF
UPCODING VERSUS ALL OTHERS**

	MEAN	STD DEV	t	P<
PRODUCT A SENSITIVITY				
Top 10 Upcoded DRGs	0.4894	0.4367		
All Others (25 with a score*)	0.2763	0.3351	1.558	n/s
PRODUCT A SPECIFICITY				
Top 10 Upcoded DRGs (9 with a score*)	0.3045	0.2787		
All Others (25 with a score)	0.0891	0.1038	2.265	.10
PRODUCT B SENSITIVITY				
Top 10 Upcoded DRGs	0.6115	0.2596		
All Others (25 with a score)	0.4808	0.3560	1.203	n/s
PRODUCT B SPECIFICITY				
Top 10 Upcoded DRGs	0.2819	0.1479		
All Others (36 with a score)	0.0594	0.0826	4.562	.05

*Note:

When a DRG may not have a sensitivity score: DRGs with no upcoding will not have a sensitivity score, as the sensitivity denominator, the number of upcoded cases, is zero. Fifteen of the 50 DRGs in our sample had no upcoding.

When a DRG may not have a specificity score: DRGs that had no cases flagged by the software will not have a specificity score, as the specificity denominator, the number of flagged cases, is zero. Sixteen DRGs had no cases flagged by Product A. Four DRGs had no cases flagged by Product B.

TABLE C-8					
SOFTWARE PERFORMANCE ON THE 10 MOST COMMONLY OCCURRING DRGs					
DRG (% of Medicare discharges)	Cases Reviewed (% of reviewed cases)	Product A		Product B	
		Sensitivity	Specificity	Sensitivity	Specificity
127 Heart failure & shock (6.3%)	245 (9.3%)	0%	0%	55%	5%
89 Simple pneumonia & pleurisy age >17 w/cc (4.0%)	299 (11.4%)	11%	4%	67%	8%
14 Specific cerebrovascular disorders except TIA (3.4%)	129 (4.9%)	58%	14%	58%	8%
88 Chronic obstructive pulmonary disease (3.3%)	163 (6.2%)	0%	0%	0%	0%
209 Major joint & limb reattachment procedures- lower extremity (3.2%)	47 (1.8%)	N/A*	N/A*	N/A	0%
79 Respiratory infections & inflammations age >17 w/cc (2.2%)	186 (7.1%)	95%	36%	55%	36%
174 G.I. hemorrhage w/cc (2.2%)	82 (3.1%)	25%	7%	38%	8%
182 Esophagitis, gastroent & misc digest disorders age >17 w/cc (2.1%)	105 (4.0%)	10%	5%	30%	7%
296 Nutritional & misc metabolic disorders age >17 w/cc (2.1%)	108 (4.1%)	36%	26%	50%	10%
112 Percutaneous cardiovascular procedures (2.0%)	9 (0.3%)	N/A	N/A	N/A	0%

*Note:
Sensitivity = N/A when we found no upcoded cases within a DRG, *i.e.*, the denominator in our sensitivity calculation is zero.

Specificity = N/A when the software did not flag any cases within a DRG, *i.e.*, the denominator of our specificity calculation is zero.

**TABLE C-9
SOFTWARE PERFORMANCE ON TEN MOST COMMON DRGs VERSUS
ALL OTHERS**

	MEAN	STD DEV	t	P<
PRODUCT A SENSITIVITY				
10 Most Common DRGs (8 with a score*)	0.2944	0.3314		
All Others (27 with a score)	0.3499	0.3896	0.364	n/s
PRODUCT A SPECIFICITY				
10 Most Common DRGs (9 with a score*)	0.1159	0.1327		
All Others (27 with a score)	0.1529	0.2028	0.484	n/s
PRODUCT B SENSITIVITY				
10 Most Common DRGs (8 with a score)	0.4405	0.2128		
All Others (27 with a score)	0.5412	0.3611	0.745	n/s
PRODUCT B SPECIFICITY				
10 Most Common DRGs	0.0817	0.1050		
All Others (36 with a score)	0.1150	0.1430	0.684	n/s

*Note:

When a DRG may not have a sensitivity score: DRGs with no upcoding will not have a sensitivity score, as the sensitivity denominator, the number of upcoded cases, is zero. Fifteen of the 50 DRGs in our sample had no upcoding.

When a DRG may not have a specificity score: DRGs that had no cases flagged by the software will not have a specificity score, as the specificity denominator, the number of flagged cases, is zero. Sixteen DRGs had no cases flagged by Product A. Four DRGs had no cases flagged by Product B.

APPENDIX D

Software Vendors' Comments



*The Tower at Erieview
Suite 3000
1301 East Ninth Street
Cleveland, Ohio 44114-1800*

*Phone (888) 538-4275
Fax (216) 687-1488
email: info@dhrystone-systems.com*

July 2, 1998

Via Facsimile and U.S.P.S. Express Mail

Dr. Wm. Mark Krushat, M.P.H. Sc.D.
Director, Research and Special Projects
Office of Inspector General
Office of Evaluation and Inspection
Department of Health & Human Services
Washington, D.C. 20201

Dear Dr. Krushat:

Thank you for the opportunity to provide our feedback and comments on the results of the FMAS Corporation study for the Office of the Inspector General (OIG) "to test the ability of commercial software products to identify Diagnostic [sic] Related Group upcoding in Medicare hospital bills."

This is to advise you that we have serious reservations regarding the study's methodology and major concerns related to many of the study's conclusions. We address the conclusions in the initial part of our response and turn to the analytical framework for the study in the latter part of this letter.

Agreement with Study Conclusions

Dhrystone Systems, Incorporated (DSI), a wholly owned subsidiary of ORION Consulting, Inc., volunteered to be one of the firms to participate in this study, because of the confidence that we had that our UB2000 DRG Auditing System was a powerful tool for detecting potential Diagnosis Related Group (DRG) upcoding. The study confirmed that confidence.

Dr. Wm. Mark Krushat, M.P.H., Sc.D.
July 2, 1998
Page 2

As noted in the conclusion of the study report:

Those hospitals that the software identified were twice as likely to have upcoded cases as a control group of hospitals. *This finding leads us to believe that the software could be used in an ongoing way to identify hospitals that are likely to upcode their Medicare bills.* (emphasis added)

There are various approaches as to how the software might be applied at the hospital level. For example, the HCFA might wish to use this software as a tool in its post-payment recovery efforts. Based on our results here, HCFA could use this software to identify hospitals that have previously demonstrated a tendency to upcode, and then perform focused review on cases from these hospitals prior to making payments.

The software also showed some success within the narrowly defined group of DRGs that exhibited the most frequent upcoding. *Because the software was relatively successful in identifying particular cases that were upcoded among these DRGs, its use here could be expected to yield significant returns.* (emphasis added). For post-payment recovery efforts, HCFA could opt to focus on cases in the upcoded DRGs; analogously, the software could be used prospectively to identify cases in particular DRGs for review prior to payment.

We agree with these conclusions and that the possible courses of action outlined have considerable potential for use by the Health Care Financing Administration (HCFA) in the development and implementation of a systematic national program for the detection of healthcare fraud and abuse.

Disagreement With Study Conclusions

The study also concluded that:

At the same time, our review leads us to raise caution about these products, particularly at the individual case level. While they worked well for the most frequently upcoded DRGs our review determined that these products were distinctly less successful for other DRGs. It is for this reason that we see only a limited role, as described above, for these products at the current time.

Dr. Wm. Mark Krushat, M.P.H., Sc.D.
July 2, 1998
Page 3

We disagree strongly with this conclusion. From our perspective, the study's results are impressive and provide a compelling rationale for using the UB2000 software as it is intended. That is, as a screening tool:

- to discover potentially problematic DRG claims, and;
- to narrow the focus for clinical review upon those DRG claims with the highest probability of upcoding.

If the software is used in this way, given the "specificity and sensitivity" results from the study, and based upon the OIG Audit of Medicare estimate that in FY 1997 there were \$4.1 billion in inappropriate DRG payments and that there are approximately 10 million medical DRG claims annually, it could:

- detect as much as \$2 billion to \$2.5 billion in inappropriate payments
- reduce clinical review to 18 percent of the total claims population rather than having to do a universal audit, for discovery purposes.

These potential outcomes are significant in terms of returns/results, time and costs savings. They should not be labeled or considered "modest" or "limited."

Study Methodology and Findings

The study itself had three basic design flaws which compromise many of its findings:

1. The sample was inappropriately drawn. Both the experimental group and the control group were outliers, or "extreme groups." As a result, it is impossible to generalize study findings to the entire universe or population.
2. To focus the study on the 50 most common DRGs biases the results and significantly understates the identification potential of the software. While these DRGs may be most common, they are not representative nor are they those DRGs in which upcoding is most frequent and financial consequences the greatest. UB2000 was designed to identify claims with the highest financial return, not claims volume.
3. Given the limited sample was based upon provider and the top 50 DRGs, extrapolation of "sensitivity" and "specificity" (or any other statistic) by DRG to the Medicare claim population is inappropriate. In fact, one DRG had a sample size of only nine.

Dr. Wm. Mark Krushat, M.P.H., Sc.D.
July 2, 1998
Page 4

There is also a major problem in the manner in which the study was administered. Although the study purports to have been conducted in a double blind manner, this was not the case.

Table 30 on page 95 of the FMAS report lists the eight categories of outcomes that cases were assigned to after review by the FMAS audit staff. A close review of that table reveals a total of 286 cases on which FMAS reviewers' initial coding differed from hospital coding. Upon review, in all of these cases the hospitals' coding was determined to be correct.

This increased the population of correctly coded cases by 14.6 percent. Because our software would have categorized these same cases as potentially upcoded, it also dramatically reduces the percent of cases that our software correctly identified as upcoded.

Given the foregoing, we must call the reliability and validity of the entire FMAS study into question.

Nature and Application of UB2000 Auditing System

The study's findings were also constrained because the nature and application of the UB2000 Auditing System was not taken into account in the study design.

The UB2000 system has proven highly successful in identifying DRG upcoding. It was originally developed for the commercial insurance market. As such, it is:

- focused on the general insured population, not Medicare.
- targeted, in terms of its primary edits, to identify high volume, high error, high dollar DRGs within that population.
- directed at claims and not at institutions and organizations.
- configured, prior to use, to reflect the client's specific payment methodology.

Further, the UB2000 software was not designed to be used on a "stand alone" basis to identify DRG upcoding. The nature of our software system is clearly stated in the marketing material for UB2000 which reads as follows:

UB2000 is a DRG auditing software package. The product focuses on the identification of problematic DRG claims. Once identified, the claim can be audited, and if coding errors exist, regrouped to a lower paying DRG. UB2000 identifies diagnosis sequencing errors, medical necessity issues, place of service questions, and DRG creep.

Dr. Wm. Mark Krushat, M.P.H., Sc.D.
July 2, 1998
Page 5

UB2000 provides tightly focused auditing where experience has shown that 80 percent of overpaid reimbursement dollars are recaptured from approximately 20 percent of all claims and the total savings from auditing can be as high as 3 percent of total DRG reimbursement. UB2000 will identify these claims for clinical review.

The UB2000 Auditing System was applied "as is" to the Medicare claims population. This was an unfair test.

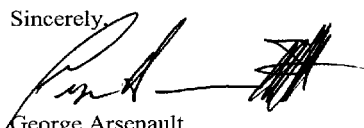
The system is easily modifiable. Had we been allowed to configure the system to the Medicare population and payment methodologies, we could have significantly improved upon its already positive performance on an "unconfigured" basis.

Conclusion and Recommendations

In conclusion, we want to commend HCFA and the OIG for the consideration given to DSI's UB2000 DRG Auditing System. We value the independent validation of the fact that our product performed as promised.

We reiterate our support for developing a strategy, based upon the OIG's suggestions and input, to use the software as a tool in an integrated national fraud and abuse detection program. In this regard, we propose that we assist HCFA in designing and implementing a "tightly targeted" audit program which will yield the greatest returns in terms of recoveries at the lowest possible costs. As part of such a program, we could configure our system and test it on a pilot basis, to demonstrate the value of taking such an approach.

Thank you for this singular opportunity to make a contribution to enhancing HCFA's fraud and abuse detection efforts. I will call you to discuss our response and to explore your interest in a revised demonstration project.

Sincerely,

George Arsenaault
Chief Executive Officer

GMA/caf

July 3, 1998



William Mark Krushat, M.P.H., Sc. D.
Director, Research and Special Projects
Office of Evaluation and Inspections
Department of Health and Human Services
Office of Inspector General
Washington, DC 20201

Dear Mark:

Thank you for the opportunity to comment on the DHHS/OIG working draft report entitled "Using Software to Detect Upcoding of Hospital Bills".

It is noted throughout the OIG report that the software that was examined including QuadraMed's DRG ✓ product, "showed modest success in identifying hospitals with a rate of upcoding..." (draft report at page ii). It is further noted at page 9 of the draft report:

The two software products that we reviewed were illustrative of what was available on the market in the Spring of 1997. We believe, however, that it is likely that the software market will continue to develop over time and that products such as these will continue to advance in sophistication and expand in their usefulness as part of a fraud detection strategy.

The DRG ✓ product continues to be upgraded and enhanced by the Compliance and Education Division of QuadraMed (Cabot Marsh). Further, DRG ✓ is now part of our Quantim FACTS suite of products (Fraud and Abuse Compliance Tool Set). The product specifications are validated by our extensive staff of Health Information Management Analysts, all of whom are credentialed experts. The upgraded and enhanced product has been named Inpatient FACTS. Inpatient FACTS includes thousands of additional microspecifications to ensure data quality for ongoing monitoring and auditing mandated under the Model Hospital Guidelines. Inpatient FACTS is utilized as a retrospective and/or concurrent monitoring and audit tool. The goal of the product is to ensure correct coding.

The addition of thousands of microspecifications into the upgraded and enhanced DRG ✓ product, Inpatient FACTS, ensures providers who license the product that their coding is consistent with true data quality. The software is objective and consistent. QuadraMed believes that in order to translate data into useful information, our transactional compliance software must continuously be enhanced and upgraded. Inpatient FACTS contains the most intensive sensitivity measures and specificity capabilities of any transactional inpatient software in the industry.

<http://www.quadramed.com>
561 Main Street • Bethlehem, PA 18018 • Phone (800) 373-5620 • Fax (610) 882-3084

Inpatient FACTS is often an integral component of a hospital's inpatient billing compliance program. At the time the study was conducted, the Model Hospital Guidelines had not been issued. Moreover, the first hospital based pneumonia DRG 079 settlement had not occurred. As a result of the DRG Creep investigations and the DRG 079 investigations, Inpatient FACTS and other software transactional products will continue to evolve and become commonplace components of compliance programs. This will ensure that providers fulfill the monitoring and auditing mandates of the Model Hospital Guidelines.

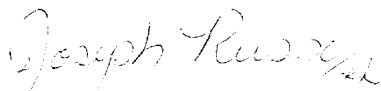
In QuadraMed's auditing and coding of over 2 million medical records, we have learned that hospitals are generally not intentionally upcoding. Error rates are generally the result of inadvertent negligence, and lack of coder education. Many coders throughout the country lack the proper credentials (not certified) and are not involved with continuous education and updates on coding changes and compliance initiatives. Moreover, insufficient and/or inconsistent physician documentation has played a part in many of the coding changes our Health Information Management experts have made during the audit process. With education becoming an integral part of provider's compliance programs, and with continuing investigations, coding error rates should be reduced.

As providers throughout the country continue to license Inpatient FACTS, we believe they will be well served by reviewing coded records concurrently (i.e., after the record is initially coded, but prior to claim submission). This ensures clean claims prior to submission to third party payors.

Software can never be utilized to replace manual review. Any software products that flag improperly coded medical records must require manual intervention prior to any coding and/or DRG changes "flagged" by software. This ensures that automated software programs are not being utilized in an inappropriate fashion. We respect the fact that manual review was mandated pursuant to the terms and conditions of the DHHS/OIG study.

It is our hope that the Final Report will make reference to the current status of our products as they have been significantly enhanced with the inclusion of thousands of additional microspecifications, focused and random sampling specifications, additional reporting features, Health Information Management expert review validation, and overall increased functionality.

Sincerely,



Joseph J. Russo, Esq.
Senior Vice President and Corporate Compliance Officer

JJR/kh

Health Information Systems
3M Health Care

100 Barnes Road
PO Box 5007
Wallingford, CT 06492-7507
203 949 0303
203 949 6331 Fax



July 9, 1998

Wm. Mark Krushat, M.P.H, Sc.D.
Director, Research and Special Projects
Office of Evaluation and Inspections
Office of Inspector General
Department of Health & Human Services
Washington, DC 20201

Dear Mr. Krushat:

We appreciate the opportunity to review the results of the *Diagnostic [sic] Related Group Code Validation Study*. Edits designed to identify potential up coding can be categorized into three broad types.

Coding Edits: These are edits that identify sequencing and other coding rule violations (e.g., when is it appropriate to code gram negative pneumonia).

Clinical Edits: These are edits that identify inconsistencies in the coded clinical information (e.g., a procedure for which there are no diagnoses present that justify the procedure).

Resource Edits: These are edits that identify inconsistencies between resource information (e.g., LOS) and the coded clinical information (an AMI patient discharged alive in one day).

The study performed focused on the identification errors that were primarily coding related. As the name Clinical Code Editor implies the CCE focuses primarily on clinical errors and to a lesser extent, resource edits. As a result, it is not surprising that the results from CCE and the two other products tested were not correlated. Unfortunately, instead of merging the results across all systems, only the results from the two correlated systems were used to select the sample for the study. Obviously, this resulted in the identification of relatively few CCE errors in the study population.

It is unfortunate that the full range of clinical errors identified by the CCE were not included in the study. Clinical errors can be used to identify significant DRG assignment problems. For example, a patient with a principal diagnosis of urethral abscess (5970) with a procedure of ureterotomy (562) will be assigned to DRG 305, Kidney, ureter & major bladder proc for non-neoplasms w/o CC, with a payment weight of 1.1695. However, the diagnosis urethral abscess does not justify the procedure ureterotomy. The procedure urethrotomy (580) would be consistent with the diagnosis urethral abscess. The change in procedure code would change the DRG to 313, Urethral procedures, age > 17 w/o CC, which has a payment weight of 0.5783.

In general, we were under the impression that the objective of the study was to identify problems that had a high probability of resulting in a DRG change without generating an excessive amount of false positives. The CCE is structured to allow the volume of cases identified as potential errors to be controlled. For example, in order to reduce the volume of cases, certain clinical edits can be restructured to only be reported if there is also an inconsistency in the use of resources. Since we assumed that false positives should be minimized, the version of CCE used in the study was very restrictive. This is evident from the study results. For the control group, which is a less biased sample than the experimental group, the CCE had a specificity of 50.0 compared to 11.5 and 5.4 for the other two systems. It would not be feasible to have a national monitoring system in which more than 9 out of 10 records selected for review were a false positive.

The CCE was designed for payors who wanted to review a small percent of claims (less than 5%) but would achieve a high return (i.e., the reduction in DRG payment would substantially exceed the cost of review). Further, the CCE was designed to encompass *all* DRGs and not to be focused on a subset of DRGs. An effective auditing system must include evaluation of all DRGs. For example, the edits conditions discussed in the study virtually ignored problems with procedure coding. Some of the largest upcoding opportunities deal with assigning patients to the "other procedure" DRGs and the use of less specific procedure codes. For example, a patient with a principal diagnosis of an open fracture of the radius (81391) with a procedure of open fracture reduction NEC (7929) is assigned to DRG 234, Other musculoskeletal system & connective tissue procedures w/o CC, which has a payment weight of 1.1126. In this example the procedure should have been open reduction of fracture of the radius or ulna (7922) which would result in DRG 224, Shoulder, elbow or forearm procedures except major joint procedure w/o CC, which has a payment weight of 0.7466.

Great caution must be used if the focus of review is only on selected DRGs. For most of the focused DRGs in the study, the DRG payment weight has been quite stable. For example, the DRG payment weight for DRG 14 and 15 are as follows.

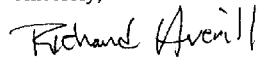
DRG	95	96	97	98	
14	1.1956	1.2065	1.1999	1.1889	CVA
15	0.6909	0.7227	0.7321	0.7241	TIA

If there was widespread coding of TIAs as CVAs, then the payment weight for the CVA DRG should have shown a significant drop. Despite the fact that there is a general lowering of payment weights over time because the DRG weights are not normalized to 1.0 each year, the payment weight for CVA has not shown any systematic decline. This simply points out the danger of having too narrow focus on the review process.

As a final point, it was disappointing that the study did not do a more specific job of quantifying undercoding. While there is certainly a significant amount of wrong coding occurring, the relative magnitude of upcoding versus downcoding is not well quantified. Further, it would have been useful to have a cost benefit analysis of the review process performed. In otherwords, would the net changes in payment justify the cost of performing the reviews?

Because of the current concern regarding compliance, 3M will be releasing a comprehensive new product called Audit Expert this fall. Audit Expert contains a complete set of clinical, coding and resource edits and allows the user to increase or decrease the sensitivity and specificity of the system to suit their needs. We would welcome the opportunity to discuss this new product with you.

Sincerely,



Richard Averill
Research Director

RFA:mm

APPENDIX E

Notes

1. Department of Health and Human Services, Health Care Financing Administration, Office of the Actuary. Personal communication April, 1998.
2. Department of Health and Human Services, Office of Inspector General, *Report on the Financial Statement Audit of the Health Care Financing Administration for Fiscal Year 1997*, A-01-97-00520, May 1998.
3. Department of Health and Human Services, Office of Inspector General, *National DRG Validation Study Special Report on Coding Accuracy*, OAI-12-88-01010, February 1988.
4. Department of Health and Human Services, Office of Inspector General, *National DRG Validation Study Update: Summary Report*, OEI-12-89-00190, August 1992.
5. The main issues of concern for those declining to participate fell into three categories: 1) the size of our test (5-10 million claims records) was too large; 2) vendors were uncertain about how OIG would utilize the results of the test; and 3) OIG's desired layout of the output was not clear enough.
6. We used the Medicare Provider Analysis and Review (MEDPAR) file as input for the software products. This file contains diagnostic, billing, and beneficiary demographic data for each stay in an inpatient hospital by a Medicare beneficiary. Our test ran approximately 6 million MEDPAR records through each software product.
7. The fact that output from one vendor's software differed significantly and that we decided not to test it is in no way a reflection on the potential merit of that software.
8. Although not the purpose of this evaluation, we also kept track of cases that were undercoded (*i.e.*, cases in which the hospital billed for a less expensive DRG than it should have). Our review found that out of 2,622 cases, 124 cases (4.73 percent) were undercoded while 254 cases (9.69 percent) were upcoded.
9. The 10 most frequent DRGs in Medicare comprise a higher percentage of the discharges in our sample compared to the all Medicare discharges (13 percent versus 10 percent) due to our sampling strategy. We sampled only among the top 50 most common DRGs.
10. FMAS Corporation. 11300 Rockville Pike. Rockville, MD 20852.
11. World Development Group, Incorporated. 5101 River Road, Suite 1913. Bethesda, MD 20816-1574.
12. FMAS Corporation. 11300 Rockville Pike, Rockville, MD 20852.

13. This represents about 6 million admissions. We used the Medicare Provider Analysis and Review (MEDPAR) file as input.

14. The fact that output from one product differed significantly and that we decided not to test it is in no way a reflection on the potential merit of that product.

15. SAS Institute, Inc. SAS Campus Drive, Cary, NC 27513.