2010 IEEE International Conference on Data Mining Workshops

Unsupervised DRG Upcoding Detection in Healthcare Databases

Wei Luo and Marcus Gallagher
School of Information Technology & Electrical Engineering
The University of Queensland
Brisbane, Australia 4072
Email:{luo, marcusg}@itee.uq.edu.au

***Abstract*—Diagnosis Related Group (DRG) upcoding is an anomaly in healthcare data that costs hundreds of millions of dollars in many developed countries. DRG upcoding is typically detected through resource intensive auditing. As supervised modeling of DRG upcoding is severely constrained by scope and timeliness of past audit data, we propose in this paper an unsupervised algorithm to filter data for potential identification of DRG upcoding. The algorithm has been applied to a hip replacement/revision dataset and a heart-attack dataset. The results are consistent with the assumptions held by domain experts.**

***Keywords*-DRG upcoding; decision tree; healthcare data; Fisher's exact test**

I. Introduction

Rising healthcare costs are an imminent issue in many developed countries. Data mining techniques can be used to better understand our healthcare systems and suggest ways to control the costs. This paper reports a data-mining application in detecting DRG upcoding, a costly data anomaly in healthcare databases.

The casemix model based on DRG has been used by many countries to fund public hospitals. A hospital in the casemix model—instead of being paid the actual cost of treating each *individual* patient—is paid the *average* cost of patients in a particular **Diagnosis Related Group** (DRG). DRGs are designed so that all patient episodes in a DRG consume similar resources—the intra-group variance is minimized. Suppose a hospital has treated in total 300 patients in DRG `O01` (Caesarean Delivery) in 2009. As the average cost of caesarean delivery is $\$13,639$, the hospital should have been paid $300 \times \$13,639$ at the beginning of 2010.

Figure 1 shows the process leading to the assignment of a DRG to a patient episode. In Step (2) of the process, a clinical coder converts the patient's medical chart into a sequence of diagnosis and procedure codes. Being a manual process that requires familiarity with medical terminology and considerable experience from the clinical coder, this coding step may introduce errors. One particularly expensive type of errors is called **DRG upcoding**. DRG upcoding refers to coding errors that result in shift of a patient episode into a DRG with higher reimbursement [8]. While DRG upcoding by private providers can be intentional, DRG upcoding in a public hospital system is most likely due to unintentional errors such as misspecification by the doctor or misunderstanding by the coder.

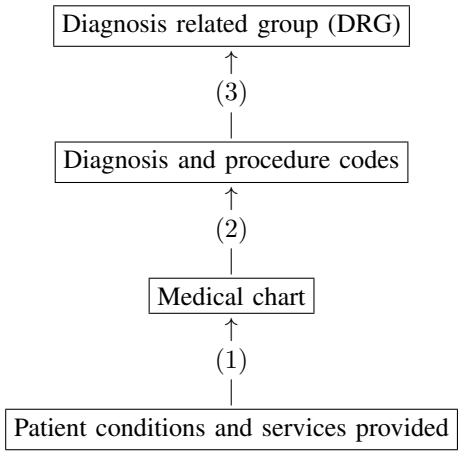

Figure 1. Flow of cost information for a typical patient episode. (1) Treating doctors hand-write diagnoses, comorbidities, complications, services, and procedures on the patient's medical chart. (2) Based on the medical chart, a clinical coder assigns a sequence of diagnosis and procedure codes. This step is where DRG upcoding is often introduced. (3) A piece of software called **grouper** maps the sequence of codes to a DRG according to a pre-programmed set of rules. The resulting DRG determines the amount of reimbursement the treating hospital will receive for the patient episode.

As DRG upcoding incurs huge cost to governments [9], [7], effective methods to detect DRG upcoding are of considerable value. Currently, there are two main ways to detect potential DRG upcoding:

1) auditing by recoding the original medical charts, and
2) comparison with historical claim data to detect increased percentage of higher-cost DRGs (e.g., [8]).

Code audit is the most reliable way to detect a DRG upcoding. In code audit, experienced health-information managers recode the original medical chart and then compare the new codes to the codes originally submitted by the hospital. The process is clearly resource intensive. Therefore the scope and frequency of auditing are severely constrained by the available resources. In reality, auditing may be done only once every ten years, and typically only fewer than 100 charts are reviewed for a hospital treating as many as $80,000$ patients a year. Year-by-year comparison is also unsatisfying

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.108

with confounding factors changing over time—such as the ageing population and unhealthy lifestyle changes in many developed countries.

In [6], a supervised model has been proposed for computing the probability that a patient episode is coded incorrectly. In this model, a training set was obtained through previous audits so that each episode was labeled to indicate whether coding error is present. The training set was used to estimate the regression parameters of a hierarchical Bayesian model. The model was then used to infer the probability that an episode contains coding error. As past audit data is often limited in quantity and scope, and is often out of date, this supervised approach has limitation in where it can be applied.

In this paper, we propose an unsupervised way to detect potential hospital-wide DRG upcoding. The basic idea behind our approach is simple. In the absence of audit data, we use data from several hospitals. We divide hospitals into a training set and a test set in the leave-one-out fashion. Records from the training set are used to define DRG-homogeneous subgroups. Then within each DRG-homogeneous subgroup, records from the test set are compared against ones from the training set. Such subgrouping properly adjusts for variation of patient population so that DRG upcoding can be identified. Although being simple, the method is effective as demonstrated by applications to real datasets. Moreover, the method requires only DRG and code data—which is readily available in every claim database—and hence has wide applicability.

The paper is organized as follows. Section II introduces relevant definitions and previous work in DRG upcoding. Section III explains the algorithm for detecting DRG upcoding. The algorithm is applied to a hip replacement/revision dataset and a heart-attack dataset in Section IV. Section V concludes the paper.

II. ICD CODES AND DRGS

In this section, we briefly review the process by which DRGs are assigned. To be concrete, we use data and examples from Australia. Nevertheless, the method we propose can be used on data from most countries with casemix funding models.

In Australia, diagnoses are coded with the **International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification**(ICD-10-AM); procedures (interventions) are coded with the **Australian Classification of Health Interventions** (ACHI). They are updated biennially and the current edition is ICD-10-AM/ACHI/ACS Sixth Edition [4]. Table I shows some example diagnosis codes in ICD-10-AM. Table II shows some example procedure codes in ACHI.

Each patient episode generates a medical chart. A clinical coder then converts the information written in the medical chart into a sequence of diagnosis codes and a sequence of procedure codes—Step (2) in Figure 1. Table III shows a sequence of diagnosis codes and procedure codes for a hypothetical hip-revision patient episode. The first diagnosis and the first procedure in the sequences are called **principal diagnosis** and **principal procedure**, respectively. They are often critical information that affects the DRG of a patient episode. In Table III, the principal diagnosis is `S72.9` and the principal procedure is `47528-01`.

Table I
AN EXAMPLE ICD-10-AM SEGMENT RELATED TO DISORDERS OF URINARY SYSTEM

Code	Description
N39	**Other disorders of urinary system**
N39.0	Urinary tract infection, site not specified
N39.1	Persistent proteinuria, unspecified
N39.2	Orthostatic proteinuria, unspecified
N39.3	Stress incontinence
N39.4	Other specified urinary incontinence
N39.8	Other specified disorders of urinary system
N39.81	Loin pain/haematuria syndrome
N39.88	Other specified disorders of urinary system
N39.9	Disorder of urinary system, unspecified

Table II
EXAMPLE PROCEDURE CODES FROM ACHI RELATED TO HIP REVISION OR REPLACEMENT

Code	Description
49315-00	Partial arthroplasty of hip
49318-00	Total arthroplasty of hip, unilateral
49319-00	Total arthroplasty of hip, bilateral
49324-00	Revision of total arthroplasty of hip
49346-00	Revision of partial arthroplasty of hip

Table III
DIAGNOSIS AND PROCEDURE CODE SEQUENCES FOR A HYPOTHETICAL PATIENT EPISODE

Code	Description
Diagnoses	
S72.9	Fracture of femur, part unspecified
W19	unspecified fall
Y92.22	Health service area
U73.9	Unspecified activity
A49.0	Staphylococcal infection, unspecified
E87.7	Fluid overload
I10	Essential (primary) hypertension
U73.9	Unspecified activity
Procedures	
47528-01	Open reduction of fracture of femur w. internal fixation
49324-00	Revision of total arthroplasty of hip
60506-00	Fluoroscopy in conjunction with surgical procedure
13706-02	Transfusion of packed cells
95550-00	Allied health intervention, dietetics

After the two code sequences are generated, a program called **grouper** assigns a DRG to the patient episode according to the code sequences, discharge status, and the patient's age and gender—Step (3) in Figure 1. In other words, the grouper is a function G such that

$$\text{DRG} = G(\text{age}, \text{sex}, \text{discharge}, \text{diagnoses}, \text{procedures}). \quad (1)$$

The DRG classification used in Australia is called **Australian Refined Diagnosis Related Groups** (AR-DRG). The system was originally derived from the DRG system used in the United States. The latest edition is AR-DRG Version 6.0, but it is AR-DRG Version 5.1 that is currently the most widely used in Australian hospitals. In AR-DRG classification, a patient episode is first assigned an **Adjacent DRG** (ADRG). Each ADRG has up to four levels of resource consumption, denoted by letters A, B, C, and D (see Table IV). For example, in AR-DRG Version 5.1, Adjacent

Table IV
ADRG SPLIT INDICATOR (AR-DRG VERSION 5.1)

Split Indicator	Interpretation within Adjacent DRG
A	highest consumption of resources
B	second highest consumption of resources
C	third highest consumption of resources
D	fourth highest consumption of resources

DRG I03 (Hip Replacement/Revision) has three levels of resource consumption, which correspond to three DRGs: I03A (Hip Revision With Catastrophic Complication and/or Comorbidity), I03B (Hip Replacement With Catastrophic Complication and/or Comorbidity + Hip Revision Without Catastrophic Complication and/or Comorbidity), and I03C (Hip Replacement Without Catastrophic Complication and/or Comorbidity). We can see that the presence of **Complication and/or Comorbidity** (CC) affects the level of resource consumption.

In the context of AR-DRG, DRG upcoding means coding errors that shift a patient episode into a higher cost DRG within an Adjacent DRG. For example, DRG upcoding can shift a hip-replacement episode in DRG I03C into DRG I03B. In year 2007-2008, average cost per episode in I03C is AU\$16,456 whereas average cost per episode in I03B is AU\$21,148 [5] (also see Table VII).

III. DRG UPCODING DETECTION

For public hospitals, we assume that DRG upcoding is unintentional. Therefore among multiple hospitals, it is unlikely that all hospitals made the same type of unintentional error. If a hospital has DRG upcoding, then a hospital's DRG distribution can be drastically different from all other hospitals. Our hypothesis is that this difference can be detected using leave-one-out cross validation. Therefore, we can use data from multiple hospitals to compensate for the absence of historical audit data.

A. General strategy

Formally, we have a collection of hospitals $\mathcal{H}$. We assume that one hospital $H \in \mathcal{H}$ has DRG upcoding. The mapping from codes to DRGs is realized with the grouping algorithm—the function $G(\cdot)$ in Equation (1). Let $G_H(\cdot)$ be the restriction of function $G(\cdot)$ on data from H and g_A be a high-cost DRG within an Adjacent DRG g. Then $(G_H)^{-1}(g_A)$, which contains all sequences of codes from H that map to g_A, should be different from $(G_{\mathcal{H}\setminus H})^{-1}(g_A)$. Unfortunately we cannot directly compare $(G_H)^{-1}(g_A)$ and $(G_{\mathcal{H}\setminus H})^{-1}(g_A)$, as they may be different even when hospital H has no DRG upcoding, as H may have a different patient population. For example, a hospital situated at a holiday spot may admit more senior and hence often more severe patients. Therefore, we have to divide all patient records into homogeneous subgroups so that the variation in patient populations is factored out before any meaningful comparison can be carried out.

To recap, we adopt the following strategy to detect DRG upcoding of a hospital H.

Step 1: Divide all cases into homogeneous subgroups.
Step 2: For each subgroup, test whether hospital H has a different DRG distribution from $\mathcal{H} \setminus H$.

As we do not know which hospital may have DRG upcoding, a cross-validation approach is used. We go through each candidate hospital $H \in \mathcal{H}$. Each time we assume no DRG upcoding for hospitals in $\mathcal{H} \setminus H$. In Step 1, the data from $\mathcal{H} \setminus H$ is used to train the rules for partitioning the space of code sequences. In Step 2, the data from H is tested against data from $\mathcal{H} \setminus H$. Figure 2 describes the algorithm with more details. Line 3 and Line 5 need further explanation and are

```
1:  for all H ∈ H do
2:      Split the data D into D_{H\H} and D_H.
3:      With D_{H\H}, derive a set of rules
        R ≜ {R_1, R_2, ..., R_n} that segments
        D_{H\H} into homogeneous subgroups
        {R_1(D_{H\H}), R_2(D_{H\H}), ..., R_n(D_{H\H})}.
4:      for all R_i ∈ R do
5:          Test the null hypothesis that R_i(D_{H\H}) and
            R_i(D_H) have the same DRG distribution.
6:          if the null hypothesis is rejected. then
7:              return H and R_i for further examination.
8:          end if
9:      end for
10: end for
```

Figure 2. Algorithm for detecting DRG upcoding.

described in Sections III-B and III-C respectively.

B. Segment patient groups

Let D be the coding data of all hospitals in $\mathcal{H}$ and H be a hospital to screen for DRG upcoding. We use D_H to denote the subset of D for H and use $D_{\mathcal{H}\setminus H}$ to denote the subset of D for all the remaining hospitals. To segment $D_{\mathcal{H}\setminus H}$ into homogeneous subgroups, we build a classification tree T that classifies DRGs based on diagnosis and procedure codes. The choice of using classification tree has the following justifications.

1) Apart from the diagnosis and procedure codes, DRG is often the only additional information available in a claim database. A classification tree with DRG as the target variable naturally incorporates DRG information into the partitioning of data.
2) In a resulting tree T, a path from the root to a leaf node N corresponds to a rule R_N which is conjunctive of multiple testing conditions. This guarantees the interpretability of the subgroups.
3) The stopping criteria of a classification-tree learning algorithm often imply the subgroups' being homogeneous.

Stopping criteria of decision tree learning: Decision tree learning can be understood as a process that repeatedly splits the leaf nodes that are "impure". For the simplicity of discussion, we shall assume that the Gini index is used for the impurity measure. For a set S of instances with K classes, the **Gini index** of S is defined to be $\sum_{k=1}^{K} \hat{p}_k(1-\hat{p}_k)$, where $\hat{p}_k$ is the percentage of instances in S that belongs to the kth class (see for example [3, Section 9.2.3]). Hence the smaller the Gini index of S is, the more pure (or homogeneous) S is. In the extreme case where every instance in S belongs to the same class, the Gini index of S is 0.

There are two reasons that the splitting process stops at a node N.

1) The node is "pure" (the impurity measure is low).
2) No other variable can split the node to significantly decrease the impurity measure.

In the first case, we have obtained a homogeneous node. In the second case, we may not have a homogeneous node, but conditional on the subset, other variables are rather independent of the target variable (DRGs in our case).

We use the standard decision tree implementation in [10]. In constructing the decision tree, for the subsequence statistical test to work, we set the minimum bucket size—the number of training instances—for a terminal node to 30. A smaller number may lead to overfitting. More details of the process in shown in Figure 3.

To demonstrate, we used a training set $D_{\mathcal{H}\setminus H}$ of size 631 in the ADRG `I03` (Hip Replacement/Revision) to learn a decision tree. One leaf node N has a bucket containing 114 instances. The rule R_N for N is

```
PRINCIPAL_PROCEDURE in (30023-00,
30241-00, 47048-00, 47516-00,
47519-00, 47522-00, 49330-00,
95550-00)
and [N39.0] == 1.
```

This is a group of patient episodes with unspecified urinary tract infection. The distribution of DRGs in $R_N(D_{\mathcal{H}\setminus H})$ is shown in Table V. The node is apparently homogeneous in that all 114 records in $R_N(D_{\mathcal{H}\setminus H})$ have the DRG `I03B`.

1: Let $D_{\mathcal{H}\setminus H}$ be the design matrix where each row corresponds to a patient episode and each column indicating the presence or absence of a code for the episode.
2: Let $\mathbf{y}_{\mathcal{H}\setminus H}$ be a vector where each element indicating the DRG split indicator for a corresponding row in $D_{\mathcal{H}\setminus H}$.
3: Train a classification tree T that uses $D_{\mathcal{H}\setminus H}$ to predict $\mathbf{y}_{\mathcal{H}\setminus H}$.
4: For each leaf node N of T, let R_N be the decision rule prescribed by the path from the root node of T to node N.
5: Let $\mathcal{R} \leftarrow \{R_N : N \text{ is a leaf node in } T\}$.

Figure 3. Algorithm for generating the rule set for homogeneous data segmentation.

Table V
DRG DISTRIBUTION OF AN EXAMPLE LEAF NODE N.

DRG	I03A	I03B	I03C
Numbers	0	114	0
Proportion	0.0	1.0	0.0

C. Group-wise comparison

We have described how Line 3 in Figure 2 is implemented. This section will cover the implementation of Line 5.

We adopt the standard way of comparing two multinomial distributions: an independence test. For a leaf node N, a cross classification table (see Table VI) can be constructed out of $R_N(D_{\mathcal{H}\setminus H})$ and $R_N(D_H)$. The table has the counts for different combinations of DRG split indictors on the one hand and sources of data on the other hand. For example, c_A^1 is the number of episodes from $R_N(D_{\mathcal{H}\setminus H})$ that belong to the DRG with split indicator A. More specifically, on Line 5 in Figure 2, we test the null hypothesis that Table VI has been generated by two independent multinomial categorical variables. The rejection of the null hypothesis indicates two different DRG distributions in $R_N(D_{\mathcal{H}\setminus H})$ and $R_N(D_H)$.

Table VI
SYMBOLIC REPRESENTATION OF A CROSS-CLASSIFICATION TABLE AT A LEAF NODE.

	ADRG Split Indicator			
	A	B	C	D
$R_N(D_{\mathcal{H}\setminus H})$	c_A^1	c_B^1	c_C^1	c_D^1
$R_N(D_H)$	c_A^2	c_B^2	c_C^2	c_D^2

There is more than one statistical test for independence. We use Fisher's exact test (see for example [2]). Fisher's

exact test can handle cross-classification tables containing cells of small counts, which is often the case for tables generated from patient subgroups.

IV. Applications to real datasets

We applied the above algorithm to a variety of datasets from 11 public hospitals across an Australian state. For this paper, we show the applications with two datasets. The first dataset consists of ICD codes and DRGs for patient episodes in ADRG I03 (Hip Revision or Replacement). The second dataset consists of ICD codes and DRGs for patient episodes in ADRG F60 (Circulatory Disorders with AMI without Invasive Cardiac Investigative Procedure). Due to the sensitive nature of the data, we number the hospitals from H_1 to H_{11}.

A. Adjacent DRG I03

In Australia, Adjacent DRG I03 (Hip Revision Or Replacement) has been ranked the 6th highest cost ADRG in year 2007-2008, with a total cost of AU$262,716,225 [5, Table 19]. ADRG I03 has 3 DRGs:

1) I03A (Hip Revision With Catastrophic or Severe CC).
2) I03B (Hip Revision With Catastrophic or Severe CC or Hip Revision Without Catastrophic or Severe CC).
3) I03C (Hip Replacement Without Catastrophic CC).

Table VII shows the different costs of the three DRGs, as reported in [5]. Due to the differences among three DRGs, DRG upcoding in this ADRG has tangible financial consequences.

Table VII
Costs of three DRGs in ADRG I03.

DRGs	cost per separation (AU$)
I03A	28,107
I03B	21,148
I03C	16,456

In this application, we retrieved a whole year's code and DRG data from the 11 public hospitals. This results in 2556 records. The decision tree partitions the records into 11 subgroups (11 terminal nodes). Only one subgroup and one hospital (H_1) has been identified by the algorithm. The identified subgroup is defined by the rule

```
PRINCIPAL_PROCEDURE in (47930-01,
49318-00, 90607-00, 92514-30) and
[Y92.22]==0 and [N39.0]==0 and
[J98.1]==0
```

The subgroup contains 1189 patient episodes without urinary tract infection or pulmonary collapse. A cross-classification table for H_1 and other hospitals is shown in Table VIII. Potential upcoding may exist for H_1 as 25% of all 59 patient episodes have the DRG split indictor B; in contrast, only 12% of all 1030 patient episodes for $\{H_2, H_3, \ldots, H_{11}\}$ have the DRG split indictor B. With Fisher's exact test, the independent hypothesis was rejected with p-value 0.008. The finding has been presented to health information managers. The users indicated that the finding confirms some assumption held by them and merits further investigation.

Table VIII
Cross-classification table for the subgroup PRINCIPAL_PROCEDURE in (47930-01, 49318-00, 90607-00, 92514-30) and [Y92.22]==0 and [N39.0]==0 and [J98.1]==0.

	I03A	I03B	I03C
H_1	0	15	44
Other hospitals	0	135	995

B. Adjacent DRG F60

In the second application, we extracted data from Adjacent DRG F60 (Circulatory Disorders With AMI Without Invasive Cardiac Investigative Procedure). **Acute myocardial infarction** (AMI) is also known as heart attack. Compared to ADRG I03, ADRG F60 consumes much less resources. It ranks only 27th in terms of the total cost in Australia [5]. ADRG F60 contains 3 DRGs:

1) F60A(Crc Dsrd+Ami-Inva Inve Pr With Catastrophic or Severe CC)
2) F60B (Crc Dsrd+Ami-Inva Inve Pr Without Catastrophic or Severe CC)
3) F60C (Crc Dsrd+Ami-Inva Inve Pr, Died)

Due to the special nature of DRG F60C, we remove from our data all records in that DRG. For the same set of hospitals in the previous application, the algorithm did not find any potential DRG upcoding. As the number of patients with heart diseases is much larger compared to the number of patients who need hip replacement, a clinical coder should be more experienced in coding episodes for the F60 ADRG. The above result may reflect uniform coding practice due to the sheer volume of cases.

V. Conclusion

This paper proposed an algorithm to filter for DRG-upcoding in the absence of historical audit data. The applications to an orthopedic dataset and an AMI dataset have demonstrated the effectiveness of the algorithm. This work shows potential value of data mining techniques in healthcare cost control, an ever more important issue facing our society.

We have used decision trees to partition data into homogeneous subgroups. A future work is to explore alternative ways to partition code spaces. In addition, when information other than codes and DRGs is available, the additional information should be used in the detection process. We are working with data managers to evaluate additional information that may increase the sensitivity and precision of the detection algorithm.

ACKNOWLEDGMENT

This work is supported by an Australian Research Council Linkage Grant (LP 0776417) and the Queensland Health Patient Safety and Quality Improvement Service. We thank Di O'Kane and Col Roberts from Queensland Health for providing access to the data and comments on the results; we also thank anonymous ICDM reviewers for helpful comments that improved the manuscript.

REFERENCES

[1] A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, 1992.

[2] Alan Agresti. *Categorical Data Analysis*. Wiley, 2 edition, 2002.

[3] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. 2009.

[4] National Center for Classification in Health. ICD-10-AM. http://nis-web.fhs.usyd.edu.au/ncch_new/2.aspx. [Online; accessed 25-June-2010].

[5] Commonwealth Department of Health and Ageing. Round 12 (2007-08) Cost Report. Technical report, the Commonwealth Department of Health and Ageing, Australia, www.health.gov.au/casemix, September 2009.

[6] M.A. Rosenberg, D.G. Fryback, and D.A. Katz. A statistical model to detect DRG upcoding. *Health Services and Outcomes Research Methodology*, 1(3):233–252, 2000.

[7] T. Schönfelder, S. Balázs, and J. Klewer. Kosten aufgrund von DRG-Upcoding durch die Einführung der Diagnosis Related Groups in Deutschland. *Heilberufe*, 61:77–81, 2009.

[8] E. Silverman and J. Skinner. Medicare upcoding and hospital ownership. *Journal of Health Economics*, 23(2):369–389, 2004.

[9] P.J.M. Steinbusch, J.B. Oostenbrink, J.J. Zuurbier, and F.J.M. Schaepkens. The risk of upcoding in casemix systems: A comparative study. *Health Policy*, 81(2-3):289–299, 2007.

[10] Terry M Therneau and Beth Atkinson. R port by Brian Ripley. *rpart: Recursive Partitioning*, 2010. R package version 3.1-46.