20 Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data

Christopher J. Matheus and Gregory Piatetsky-Shapiro
GTE Laboratories Incorporated

Dwight McNeill
GTE Service Corporation

Information by itself is a pretty thin meal, if not mixed with other ingredients.
– *Internet quote*

Abstract

One of the most promising areas in Knowledge Discovery in Databases is the automatic analysis of deviations. Success in this task hinges on the ability to identify a few important and relevant events among the multitude of potentially interesting deviations. In this chapter we present our approach to determining the interestingness of a deviation via the potential benefit from a relevant action. This approach has been implemented in the Key Findings Reporter (`KEFIR`), a system for discovering and explaining "key findings" in large, changing databases, currently being applied to the analysis of healthcare data. The system performs an automatic drill-down through data along multiple dimensions to determine the most interesting deviations of specific quantitative measures relative to their previous and expected values. It explains "key" deviations through their relationship to other deviations in the data, and, where appropriate, generates recommendations for actions in response to these deviations. `KEFIR` uses Mosaic, a WWW browser, to present its findings in a hypertext report, using natural language and business graphics.

20.1 Introduction

Increasingly, databases are being used to record all kinds of business transactions. The timely analysis of these databases is highly desirable and may often provide a competitive advantage. As these databases grow larger – with gigabyte sizes quite common – they are overwhelming the traditional query and report based methods of data analysis. This has led to the development of systems for automated discovery in databases (Piatetsky-Shapiro & Frawley 1991, Piatetsky-Shapiro 1993, Fayyad & Uthurusamy 1994).

While many different discovery methods (Matheus et al. 1993) have been tried, some of the most successful business applications – such as the Spotlight (Anand and Kahn 1992) and CoverStory (Schmitz, Armstrong, and Little 1990) systems for supermarket sales analysis – have been based on the detection of deviations, i.e. significant differences between measured values and corresponding references such as previous values or normative values.

Deviations are powerful because they provide a simple way of identifying interesting patterns in the data. We have studied many knowledge discovery algorithms with potential for identifying vast numbers of significant patterns from data, but most of these are unable to determine when a pattern is truly interesting to the user (Matheus et al. 1993). With deviations we have a simple way to identify things that differ from our expectations – since they differ from what we expect, they are by definition interesting at least to some degree.

`KEFIR` is a system for discovering, explaining, and reporting on key deviations in large databases. It performs an automatic drill-down through data along multiple dimensions to determine the most interesting deviations of specific quantitative measures relative to norms and previous values. It then explains "key" deviations through their relationships to other deviations in the data, and, where appropriate, generates simple recommendations for actions in response to deviations. The results are compiled into a written report in the form of a hypertext document, delivered using the World Wide Web (WWW) (Berners-Lee et al. 1992).

In this chapter we describe `Health-KEFIR` which is an application of `KEFIR` to the healthcare domain. In particular, we will concentrate on the problem of how to decide which deviations are the most interesting. We will argue that to properly judge the interestingness of a deviation, one should examine deviations between an observed value and *all* relevant reference values, that is all previous values, and all relevant normative values. Such a set of related (current, previous, and normative) values of a given measure is called a *finding*.

Central to `KEFIR`'s methodology is its ability to rank deviations according to some measure of "interestingness" to the user of the system. In earlier work (Piatetsky-Shapiro

1991, Frawley, Piatetsky-Shapiro, and Matheus 1992) we examined various factors of interestingness such as novelty, utility, relevance, and statistical significance. Here, we will argue that, at least in business applications, domain-independent statistical measures of interestingness are insufficient and that the interestingness of a finding should be based on the estimated utility that could be realized by taking a specific responsive action.

In the following section we describe the process of data analysis of healthcare information, which sets the stage for our analysis of interestingness. We then describe the complete `KEFIR` system, and offer ideas for future work in this area.

20.2 Healthcare Data Analysis

With the steady rise in healthcare costs and the growing urgency to control these costs, timely analysis of healthcare information has become an issue of great importance. Large corporations, hospitals, health-maintenance organizations, and insurance companies all require expert analysis of their healthcare data – an endeavor that is both time consuming and very expensive. A single report may take weeks or months to prepare and can cost tens of thousands of dollars. For large corporations, which typically order many reports for different business units, healthcare consulting costs may run into millions of dollars per year. The great time and expense of preparing a report acts as a disincentive to ordering them in many cases, thus eliminating potential savings opportunities. Even when a report is ordered, it may be incomplete, because an exhaustive search of possible findings and their explanations is simply infeasible by manual means. In our own experience, we have seen examples where human experts overlooked important findings later detected by `Health-KEFIR`.

This situation presents a real opportunity for automating data analysis and reporting systems, especially because the methods currently employed by healthcare analysts lend themselves well to automation. These methods rely on a set of relatively standard *measures* which assess various aspects of healthcare, such as cost, price, usage, and quality (e.g. `average_hospital_payments_per_capita`, `admission_rate_per_1000_people`, `cesarean_section_rate`). These measures are usually aggregate values taken over subsets of data. These subsets are created along two independent dimensions. One dimension defines the *population group* of interest, e.g. a business unit, a region, union employees, etc. The other dimension defines the *study area* (the type of medical problem to be analyzed for the selected population group): e.g. inpatient admissions, premature pregnancies, elective surgeries, etc.

A fundamental question in healthcare analysis is: For a given population group and study area, how do the current measure values compare to previous values and to normative values? If a measure for the population has changed dramatically or deviates significantly from the norm, then this is a potentially interesting finding. The actual interestingness depends on whether there are actions that can be taken in response, and on the benefits that might result. For example, a \$1,000,000 jump in payments due to an increase in the number of regular pregnancies is probably less important than a \$200,000 increase in payments due to premature deliveries, since there are no relevant actions for the first finding, while there are well-established intervention strategies for the second finding that can save a significant part of the cost *and* improve the quality of care. Thus, the *interestingness of a deviation is related to the estimated benefit achievable through available actions* – a notion which is quite generic and not limited to healthcare.

20.3 Analysis of Interestingness

In this section we present a principled approach to the analysis of interestingness.

The interestingness of a deviation, in brief, is determined by the payoff from a relevant action that can be taken in response to that deviation[1]. We note that in healthcare analysis, the interesting deviations are generally those which indicate a problem that can be corrected. In other applications deviations could lead to opportunities of other types.

In our discussion, we use D to denote a database instance and S to denote a subset of data (called a sector). A *measure*, denoted $M(S, D)$ (or simply M) is a function that returns a value when applied to a particular sector S and a database instance D. Not all measures are applicable to all sectors. We also assume, unless noted otherwise, that the "interesting" direction for each measure is up, i.e. the increase in a measure such as `payments_per_case` (while bad for GTE) is interesting for health-care analysts, since it represents an opportunity for savings.

First, we will examine the interestingness of a single deviation and then show why it is important to combine temporal and normative deviations.

20.3.1 Impact of a Deviation

The basic measure of importance for business events is their financial impact. Other measures, such as quality or customer satisfaction, can also be translated to financial terms (although difficult, it is done, for example, by lawyers). The financial impact of a deviation serves as the objective component in `KEFIR`'s measure of interestingness.

Let us begin with a concrete example. If `payments_per_case` for surgical admissions in the west region increased from \$14,818 to \$23,187 between 1992 and 1993, what is the financial impact of this change on the bottom line? To answer this question, we first

[1] This fits into a statistician's view of an optimal utility function, as defined, e.g. by (DeGroot 1970)

need to select a measure M_0 that represents the bottom line. This measure should be such that any other measure M_i can be related to M_0 via some function f_i such that $M_0 = f_i(M_i, D)$. Note that f_i would generally be a function of other measures and also of the database instance. For `Health-KEFIR` M_0 is the total healthcare payments, denoted `total_payments`.

The impact should be measured with respect to the selected population group P_0 and the top-level study area (which contains all episodes of care) A_0. The intersection of P_0 and A_0 is called the top-level sector, denoted S_0.

Formally, the *impact* of the deviation of measure M_i in sector S_j from its value $M_i(S_j, D_R)$ in the reference database D_R to its value $M_i(S_j, D_C)$ in the current (observation) database D_C, denoted $\mathit{impact}(M_i, S_j, D_C, D_R \| M_0, S_0)$, is the difference between

- the value that the bottom-line measure M_0 would have if the value of M_i for sector S_j was changed to its current value $M_i(S_j, D_C)$, while all other values would be as in D_R, and
- $M_0(S_0, D_R)$, the reference value of M_0 in sector S_0.

The equation for computing impact is:

$$\mathit{impact}(M_i, S_j, D_C, D_R \| M_0, S_0) = f_i(M_i(S_j, D_C), D_R) - M_0(S_0, D_R)$$

When the values of M_0 and S_0 are obvious they will be omitted.

For example, if `payments_per_case` in sector S_{surg} increased from \$14,818 in 1992 (D_R) to \$23,187 in 1993 (D_C), and `cases`(S_{surg}, D_{92})=149, then the impact of just that change on the bottom line measure can be computed to be (23,187 - 14,818) $\times$ 149 = \$1,246,981. The formulas for computing the impact are described in more detail in the appendix.

In this example the impact represents the potential savings that would be realized if the current value of the measure was brought back to its previous value, i.e. GTE could save \$1,246,981 if the `payments_per_case` measure for surgical admissions was brought back to its 1992 level. This approach has two major problems: (1) the old value may be an unrealistic target, and (2) the impact does not indicate the degree of control or *discretion* we might have in changing the measure. These two problems are discussed in the following sections.

20.3.2 Combining Trend and Normative Deviations

CoverStory, an earlier system for finding deviations in supermarket sales data (Schmitz, Armstrong, and Little 1990) looked at changes over time in a small number of predefined measures such as sales volume and share, and five additional causal measures (called factors in CoverStory) – distribution in store, price, store displays, feature ads, and price cuts – pre-determined by earlier market research to have effect on changes in sales volume. The interestingness of a change in a factor value was determined heuristically as:

$$Interestingness = \%Change \times FactorWeight \times MarketWeight$$

Here the *FactorWeight* is a weight on one of the five causal factors, and *MarketWeight* is the square root of the market size.

In healthcare management it is not sufficient to look only at changes over time, since the past is not always the best reference for comparison – bringing a measure to its old value may in fact be quite difficult if not impossible. Instead, we are concerned with what can be done now to reduce future costs. A more realistic target in this case is the measure's expected value. In healthcare, as in other fields, there are tables of national, regional and other norms for many key measures. These tables can be used to derive expected values for a given population (Piatetsky-Shapiro and Matheus 1994).

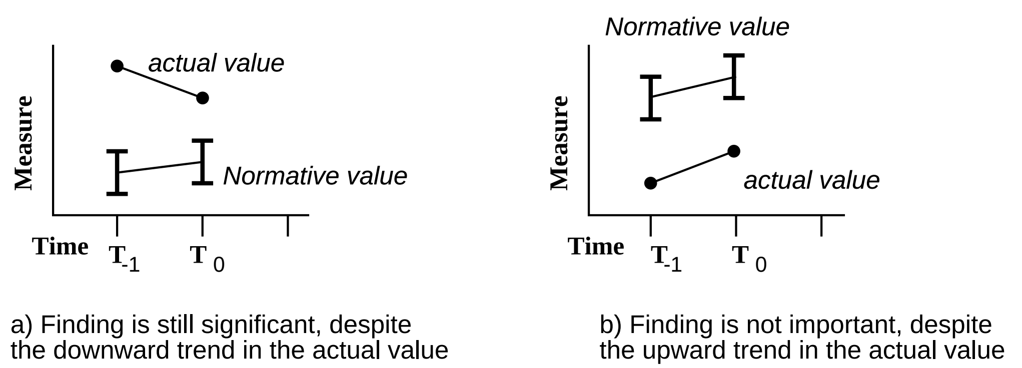


Figure 20.1
Focusing only on changes can be misleading

Figure 20.1 shows two examples of how focusing only on changes can be misleading. In Figure 20.1a, despite the downward trend in the measure, the finding is potentially interesting because the measure's value remains significantly above the norm. In Figure 20.1b, despite the upward trend in the measure, the finding is probably not as interesting because the measure's value continues to be below the norm. It is also insufficient to look only at present deviation from the norm, since this ignores the trends of the measure and of the norm.

In any case, the deviation at present reflects the potential savings which were already missed, and is only an approximation for the real measure of benefit, which is the "potential savings" achievable in the future. To determine that we need to 1) forecast the measure's future deviation from the norm, 2) translate that deviation into impact on the bottom line, and 3) determine how much of that amount we can expect to save by intervening with an appropriate action (see figure 20.2). Let us look at each of these steps in turn.

1) The Projected Deviation From the Norm: In KEFIR we forecast measure values at the end of the year by making a simple linear projection based on values from previous years. More complex forecasting strategies will be considered as additional historical data becomes available. A further refinement we are considering is knowledge-based forecasting (Lee, Oh, and Shin 1990), which could take into account domain knowledge of trends in medical inflation, normative measures, etc.

2) Translating the Projected Deviation Into Bottom Line Impact: After forecasting the future actual and normative values of the measure M in sector S at time T_1, we can compute the projected impact of this difference on the bottom line (see Appendix). When the formula relating M to the bottom-line measure uses only additions or multiplications, the impact can be simply computed by multiplying the projected deviation by the impact factor.

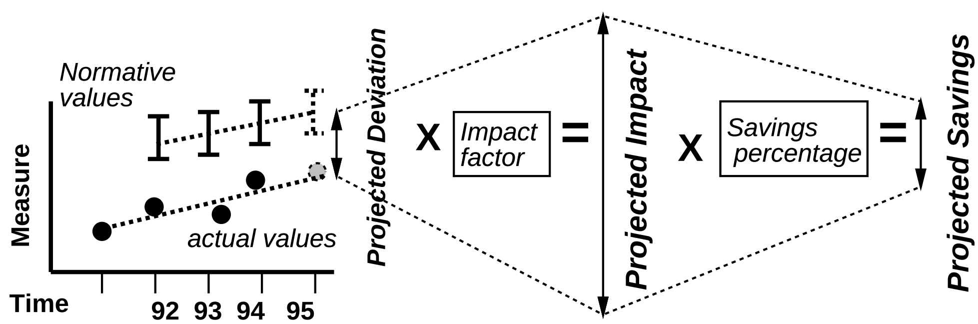


Figure 20.2
Determining future potential savings

3) Actions to reduce the deviation: The projected impact on the bottom line is the amount that can be saved if we could bring the measure's value to the projected norm.

Changing the trend of a measure, however, is possible only when relevant intervention strategies are available. For example, GTE healthcare managers have several accepted actions for containing the cost of chronic care or for reducing admissions for premature pregnancies, while they have no actions for affecting admissions for normal pregnancies.

Our healthcare domain expert has provided a number of recommended actions for various measures in different study areas. These are encoded in the system as production rules. For each rule, the expert estimated the savings percentage, i.e. the percentage of the deviation from the norm that can be saved by taking the action.

For a given finding, the system identifies matching rules and selects the rule with the highest savings percentage. It then computes the projected savings as:

$$ProjectedSavings(finding) = ProjectedImpact(finding) \times SavingsPercentage$$

Because the healthcare field is rapidly progressing, the set of available actions will be constantly changing. To adjust to that, and to explicitly account for incompleteness in the system's knowledge, a default action of simply reporting the deviation matches any finding; this can be viewed as an encoding for the likelihood that bringing the deviation to the user's attention will lead to some (unknown to the system) corrective action. The savings percentage of the default action, however, is generally low.

20.3.3 Statistical Significance

Let us further consider the example of deviations in Surgical `payments_per_case`. The significance of this deviation would be less if the million-dollar-plus increase were attributable to a single extreme case than if it were due to several dozen high-cost cases. The rationale for this reasoning is that a single extreme case is unlikely to re-occur next year, and so there is nothing to be done; several dozen high-cost cases, however, indicate a potentially correctable pattern. Formally capturing this intuition requires analysis of statistical significance.

Estimating the potential benefit of an action as a single number (e.g. estimated benefit = $567,432) has the added problem of giving a false sense of precision. Forecasting is intrinsically an imprecise science and it would be much better to give a range and a confidence level (e.g. estimated benefit is between $400,000 and $700,000 with confidence 0.9), or even a central estimate and a standard deviation.

Computing the confidence or a standard deviation requires either knowing the a priori data distribution (impossible in our application and in most real cases), or having a large set of historical data points. In our application, we have huge amounts of data, but at this writing they only go back two years, and thus we cannot make a reliable annual forecast based only on this data. The lack of historical data and the resulting lack of standard statistical measures is, unfortunately, typical for many areas of medical cost analysis today. In the meantime, we are solving the problem by using simple approaches such as discounting findings based on less than a minimum number of cases, and using heuristic rules for dealing with extreme deviations based on a small number of cases. Methods for producing better estimates, given very incomplete data, are the topic of further research.

20.4 The KEFIR System

KEFIR implements the described approach to determining interestingness. However, a complete discovery system requires much more than just detection and ordering of deviations. The deviations are intricately interconnected and frequently a high-level deviation needs to be explained via other deviations. In the area of healthcare at least, an analysis is incomplete without some recommendations for how to remedy the problems represented by the discovered deviations. Finally, the results need to be presented in a user-friendly manner.

The overall design and process flow of the system is depicted in Figure 20.3. The healthcare data, the normative data, and domain knowledge (e.g. the measures, the population groups, the study areas, and the recommendations) are the sources of input. The system calculates all deviations, uses relevant recommendations to order deviations according to their interestingness, generates explanations for the most important deviations, and compiles the results into a written report with text, tables, and charts.

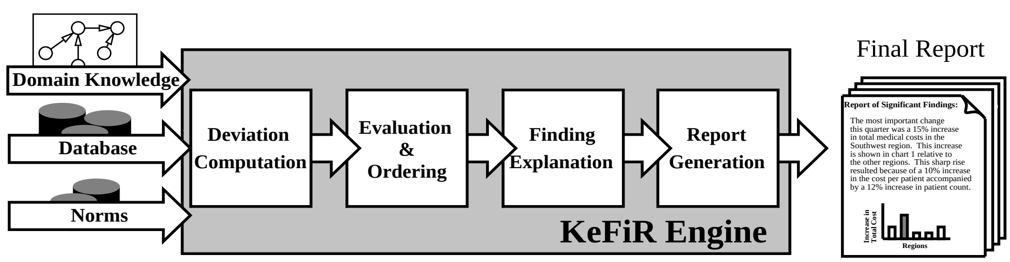


Figure 20.3
Overall design and process flow within KEFIR.

20.4.1 The Search Space

The deviation search space that KEFIR explores has three separate components (see figure 20.4). The first component is the population group of interest. The entire population is broken down into different populations subgroups, using criteria such as organizational unit, geographical region, employee status, etc.

The second component is the medical study area (corresponding to types of medical problems and treatments). The top-level study area breakdowns, such as inpatient vs outpatient, are the same for all population groups. On lower levels we have fields like DRG (Diagnostic Related Group), which is a 3-digit code for the medical problem type. Since there are almost 500 distinct DRGs and lower-level study areas for different populations will generally have only a few (but different) DRGs, it does not make sense to create sub-study areas for all 500 DRGs. Instead, KEFIR dynamically computes the top N (typically 3 to 5) DRGs with the highest payments, and limits the further analysis only to those DRGs.

The third component is the set of relevant measures. These measures are independent of the population group, but do depend on the study area. While some very generic measures, such as total payments, are defined for all study areas, most measures are meaningful only for a subset of study areas (e.g. average length of stay is only meaningful for inpatient admissions).

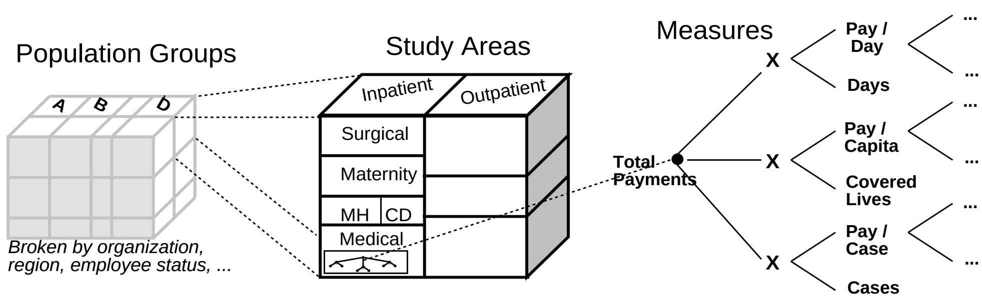


Figure 20.4
KEFIR search space of population groups, study areas, and measures

20.4.2 Deviation Detection

The deviation search space that KEFIR explores is specified by the selected population group, the study area hierarchy, and a set of measures relevant for each study area. We refer to the *intersection* of the population group and a study area as a *sector*. The "top sector" refers to the top-level ("All episodes") study area for the selected population group. KEFIR begins its analysis by evaluating the trend and normative deviations of all the measures relevant to the top sector. New sectors are then created for each of the partitions defined by all relevant categories, and deviations are calculated for each measure in each of these new sectors. This drill down into smaller and smaller sectors continues recursively until the search space is exhausted or the number of episodes in the sector becomes too small. The result of this detection process is several hundred to several thousand deviations.

Deviations are encoded in `KEFIR` within structures we call *findings*. Each finding stores information for a single measure within a single sector. Both the trend and normative deviations are stored within the finding structure. Additional information is also maintained regarding the the impact of the finding, its relation to other findings (for use in explanation), matching recommendations, references to relevant measure and sector information, and miscellaneous book-keeping data.

20.4.3 Ordering Deviations

After the deviations are calculated, they are ordered in preparation for selecting the *key findings* to include in the final report. This ranking uses the interestingness measure defined earlier in this chapter. The top N findings are marked as "key findings," with N being a user definable parameter that defaults to ten.

20.4.4 Explanation

`KEFIR` generates explanations for all its key findings. An explanation for a given finding can come from the decomposition of a formula that defines the finding's measure, or from the breakdown of the measure into its values from the sub-sectors derived from the finding's sector. The decomposition of a measure by formulas is shown in Figure 20.4. In this example, the measure `total_payments` can be decomposed by three different formulas. The factors in these formulas are drivers of the `total_payments` measure since a change in any one directly affects a change in the value of `total_payments`. Using this knowledge, we can begin to explain an observed deviation in `total_payments` by relating it to the factor most responsible.

The breakdown of a sector into sub-sectors is illustrated in Figure 20.5. The high level `inpatient` sector can be broken down into sub-sectors by several different categories. The highlighted category in this example, `admission_type`, breaks the `inpatient` sector into four disjoint sub-sectors. If a deviation is observed in a measure, such as `total_payments`, we can determine which if any of these sub-sectors is most responsible by comparing their own deviations for that measure. Although this example shows but a single breakdown, in practice there may be many levels, resulting in increasingly smaller and more homogeneous sub-sectors.

`KEFIR` explains a key finding by first evaluating all other findings affecting it through

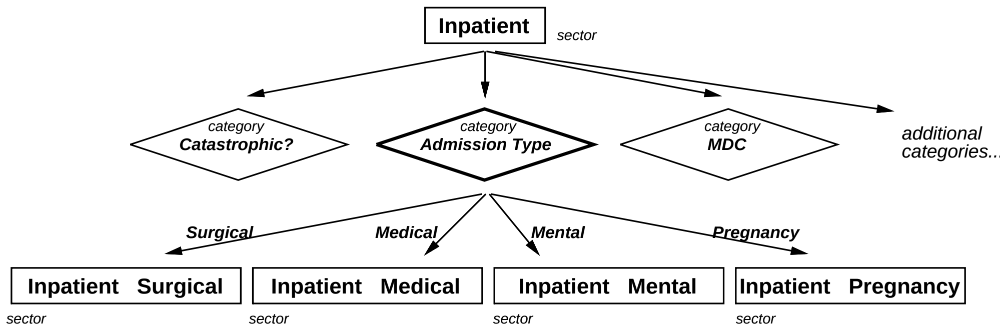


Figure 20.5
Examples of how a sector is broken down into sub-sectors using predefined categories. The resulting tree of dependent sectors is used to explain deviations by tracing the value of a measure down into the sub-sectors contributing the most to the observed deviation at the high level.

formulas or breakdowns. It then selects the one finding with the greatest influence and attempts to explain it in the same manner. This recursive process continues until there are no more interesting findings to explain, or a predefined depth (typically of three) is reached. The final result is a sequence of explanations that chain together a set of interesting findings.

20.4.5 Recommendation

The main purpose for reporting the key findings is to indicate areas where delivery of healthcare could be improved. In many cases, the information provided by a finding is sufficient for the system to automatically suggest a specific course of action for handling the problem. `Health-KEFIR` uses a set of rules to identify these situations and to generate recommended actions. The following is the content of a simple recommendation rule:

```
IF measure = admissions_per_1000  &   sector = Premature_Pregnancies &   percent_change
```

The recommendations were defined by our healthcare expert. Approximately 35 recommendations currently cover the most general healthcare areas. Refinements and additions to this knowledge base are on-going. With the recent addition of a WWW-based editor, expert users are now able to add and modify recommendations remotely.

20.4.6 Report Generation

The final output from `KEFIR` is a written report of the key findings, their explanations, and recommendations. Sentences and paragraphs are generated using simple template matching, with the option for randomized variations to produce more natural sounding text. Descriptive information relevant to the findings also appears in the report in the form of tables, bar charts, and pie charts. The report is generated in hypertext format using HTML (hyper-text markup language) to allow viewing with a WWW browser (Berners-Lee et al. 1992). When hardcopy output is desired, the HTML files are converted by a program to LaTeX, and then into postscript for printing. Figures 20.6 and 20.7 depict pages from an actual `Health-KEFIR` report.

20.5 The Implementation of KEFIR

`KEFIR` was written entirely in tcl (Ousterhout 1994) and C, making it widely portable across platforms. The system's access to data is implemented through an SQL interface which ensures portability to a wide range of database servers. We are currently running the system on a Sun SPARCstation 20 with an Informix DBMS. The design and development of `KEFIR` required approximately two full-time-employee (FTE) years. Another six FTE-months went into the knowledge engineering required to construct the knowledge base for `Health-KEFIR`. The bulk of this knowledge is represented in a collection of study areas, categories, and measures. Figure 20.8 shows parts of the structure definitions for typical instances of these objects.

`Health-KEFIR` performs its analysis on a central workstation, but it makes its results available remotely by creating a collection of HTML (hypertext markup language) and GIF (graphic interchange file) files, and serving these over the network using NCSA's httpd (hyper-text transfer protocol) server. The information manager for GTE's Managed Healthcare Program accesses the reports using a WWW client (such as NCSA's Mosaic). From there the reports can be printed for wider distribution or copied into local files for editing into specialized reports. The user may also request the generation of new reports or edit the knowledge base of recommendations through the same WWW client.

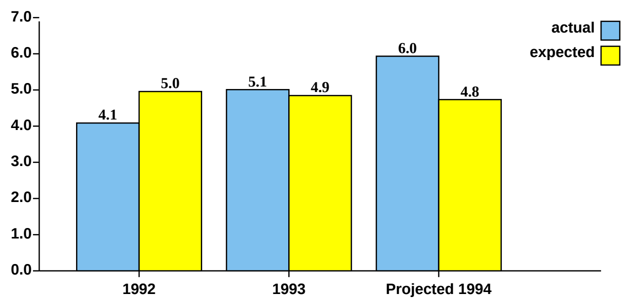
20.6 Deployment and Testing

The initial version of `Health-KEFIR` was deployed in January of 1995. Deployment to GTE's regional managers across the country is scheduled for later in 1995. Initial feedback from users has been extremely positive. In comparison to healthcare consultant reports, `Health-KEFIR` tends to identify the same general problem areas and often uncovers new, unexpected findings deemed interesting by the users. More rigorous tests will follow as new quarters of data arrive and are turned into reports.

Medical Admissions

Significant Findings

Actual and expected values for average length of stay:



Average length of stay in this study area rose 22.6 percent, from 4.1 to 5.1. Around $36,000 could have been saved if average length of stay had been equal to the expected value of 4.9. If current conditions continue into the next period, this trend will result in $263,000 of additional expenses.

Explanations: The increase in average length of stay in this study area is related to the fact that average length of stay in Medical Nervous System (MDC=01) rose 247.9 percent, from 4.4 to 15.2.

Recommendation: These admissions are for medical disorders, conditions like diabetes, heart attack, stroke. Non-medical admissions, e.g. mental health, surgical, maternity, are easier to control because there are specific care management programs. Programs that are disease or condition specific are also starting to develop. Chronic care management is probably the most important to concentrate on for medical admissions.

Projected savings: **$79,000** (based on an expected payoff of 30%).

 Medical Admissions

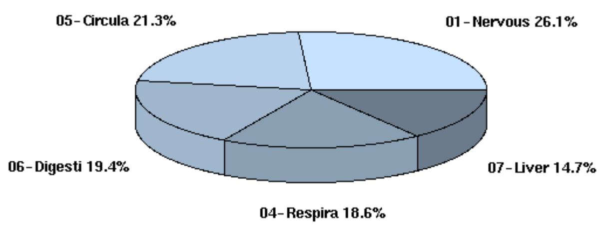
Figure 20.6
Sample output from a KEFIR report.

Total Payments Summary

Total inpatient payments in this study area fell 22.5 percent, from \$1.4 million to \$1.1 million. The fact that total inpatient payments was better than the expected value of \$1.2 million accounted for a savings of \$147,000. If current conditions continue into the next period, this trend will result in \$388,000 of savings.

The decrease in total inpatient payments in this study area is related to the fact that total inpatient payments in Medical Circulatory System (MDC=05) fell 58.7 percent, from \$370,910 to \$153,287. The observed deviation from the expected value is related to the fact that payments per day was 8.1% below the expected value (\$910.68 versus \$1,479.59).

Percent of Total Inpatient Payments by Top 5 MDC by payments



Measures Table

Measure	1992	1993	Change	1993 Norm	Diff.
Total payment	\$1,415,839	\$1,097,380	-22.5%	\$1,244,343	-11.8%
Pay/capita	\$135	\$125	-7.8%	\$141	-11.8%
Covered lives	10441	8778	-15.9%		
Episodes	299	237	-20.7%		
Average LOS	4.1	5.1	22.6%	4.9	3.4%
Pay/day	\$1,141	\$910	-20.2%	\$1,479	-38.5%
Adms/1000	28.6	27.0	-5.7%	19.5	38.6%
Readmit-30 rt	12.4	16.0	29.6%		
Mjr adv rt	3.3	2.5	-24.3%		

 Medical Admissions

Figure 20.7
Sample output from a KEFIR report (continued).

```
Category: Admission_type  name:     {admission type}  split_by: ADMTYP  values:   {value: 1
```

Figure 20.8
Samples of three of the structures used in `Health-KEFIR`. The slots shown in these examples are only those that have predefined values, i.e. they represent elements of domain knowledge. Additional slots exist for each structure, the values of which are filled in at run time.

20.7 Limitations and Extensions

The performance of `Health-KEFIR` is only as good as its domain knowledge. We are adding to its rule base to broaden and improve its recommendation capabilities. New categories and measures are being added to reflect changes in the way healthcare managers desire to aggregate and decompose information.

Currently the system only handles simple trend and normative analysis. We would like to extend its capabilities to include trend analysis over multiple periods, and add model-based comparisons. While norms are useful references for average performance, it is often desirable to set other targets for comparison. For example, rather than comparisons to the average it has been argued that comparisons should be made to a "best practice" model, i.e. comparison to a target representing an achievable level of above average performance. In some situations unusual circumstances may make even the average unachievable, in which case we might wish to set sights on some target below the norm.

20.8 Concluding Remarks

`KEFIR` represents an approach to developing complete, automated systems for identifying, explaining, and reporting on key deviations in large, changing databases. The successful implementation of `Health-KEFIR` demonstrates the merits of the system and the potential power of the proposed methodology. This technology has matured to a point where wider application to other domains is now feasible and desirable. We are currently considering an application of `KEFIR` to databases in the areas of marketing and customer analysis.

Acknowledgments: We are very grateful to Shri Goyal and Bill Griffin for their encouragement of our work on discovery in databases. Zhuma Feng made significant contributions to many parts of `KEFIR`. We thank Bernard Silver and Roland Zito-Wolf for their comments on drafts of this chapter.

Bibliography

T. Anand and G. Kahn, 1992. SPOTLIGHT: A data explanation system. In *Proc. Eighth IEEE Conference on Applied AI.*

T. J. Berners-Lee, R. Cailliau, J-F Groff, and B. Pollermann, 1992. World-Wide Web: The Information Universe, *Electronic Networking: Research, Applications and Policy*, Vol. 2 No 1, pp. 52-58 Spring Meckler Publishing, Westport, CT, USA.

M. H. DeGroot, 1970. *Optimal Statistical Decisions.* New York: McGraw-Hill.

U. Fayyad and R. Uthurusamy, eds., Proceedings of AAAI-94 Knowledge Discovery in Databases Workshop, Menlo Park, Calif.: AAAI Press Tech Report WS-94-03.

J. K. Lee, S. B. Oh, and J. C. Shin, 1990. UNIK-FCST: Knowledge-assisted adjustment of statistical forecasts. *Expert Systems with Applications*, 1(1):39–49.

C. J. Matheus, G. Piatetsky-Shapiro, and D. McNeill, 1994. An application of KEFIR to the analysis of healthcare information, In U. Fayyad and R. Uthurusamy, eds., *Proceedings of AAAI-94 Knowledge Discovery in Databases Workshop.* Menlo Park, Calif.: AAAI Press.

C. J. Matheus, P. K. Chan, and G. Piatetsky-Shapiro, 1993. Systems for knowledge discovery in databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(6):903–913, December.

D. McNeill, 1993. A comprehensive set of performance measures to evaluate managed healthcare organizations: GTE's perspective. In *National Quality Management Conference*, December.

J. Ousterhout, 1994. *Tcl and TK toolkit.* Reading, Mass.: Addison-Wesley.

G. Piatetsky-Shapiro, 1991. Discovery, analysis, and presentation of strong rules, In G. Piatetsky-Shapiro and W. J. Frawley, 1991, eds., *Knowledge Discovery in Databases.* Menlo Park, Calif.: AAAI Press.

G. Piatetsky-Shapiro and W. J. Frawley, 1991. Editors, *Knowledge Discovery in Databases.* Menlo Park, Calif.: AAAI Press.

G. Piatetsky-Shapiro and C. J. Matheus, 1994. The interestingness of deviations, In U. Fayyad and R. Uthurusamy, eds., *Proceedings of AAAI-94 Knowledge Discovery in Databases Workshop.* Menlo Park, Calif.: AAAI Press.

G. Piatetsky-Shapiro, 1993. Editor. *Proceedings of AAAI-93 Knowledge Discovery in Databases Workshop*, Menlo Park, Calif.: AAAI Press Tech Report WS-93-02.

J. Schmitz, G. Armstrong, and J. D. C. Little, 1990. CoverStory – automated news finding in marketing, In *DSS Transactions*, pages 46–54, Providence, RI. Institute of Management Sciences.

Appendix 20.A Calculating Impact of a Deviation

The bottom line measure M_0 is chosen so that all other measures can be related to it via formulas. Thus, for any measure M_i there is a function f_i such that $M_0 = f_i(M_i, D)$ and the impact of change in M_i is computed as

$$\mathit{impact}(M_i, S_j, D_C, D_R \| M_0, S_0) = f_i(M_i(S_j, D_C), D_R) - M_0(S_0, D_R) \tag{20.8.1}$$

First, let's examine the computation of impact of change in the bottom line measure M_0 for different sectors. If $S_0 = S_1 \cup S_2 \cup \ldots \cup S_k$, (where all S_j are disjoint) then we can write the old value of M_0 for top sector as

$$M_0(S_0, D_R) = M_0(S_1, D_R) + ... + M_0(S_j, D_R) + ... + M_0(S_k, D_R) \tag{20.8.2}$$

and the value $M_0^{(j)}$ that M_0 would have if M_0 would change only in S_j but not in other sectors, is

$$M_0^{(j)}(S_0, D_R) = M_0(S_1, D_R) + ... + M_0(S_j, D_C) + ... + M_0(S_k, D_R) \tag{20.8.3}$$

Subtracting these equations we get

$$\mathit{impact}(M_0, S_j, D_C, D_R) = M_0(S_j, D_C) - M_0(S_j, D_R) \tag{20.8.4}$$

i.e. the bottom-line impact of M_0 change in S_j is simply the difference between the new and the old values of M_0 in S_j. For example, if the total payments for surgical admissions changed from \$2.2 million in 1992 to \$3.2 million in 1993, the impact on the bottom line would be \$1 million.

Next, we focus on a specific sector S and examine how to compute the impact of change of a specific measure M_i (other than M_0) in just that sector.

Measures are related to M_0 by different formulas. In general, we can use equation 20.8.1 to determine the impact. However, there is a much simpler method in the important special case when these formulas have only additions and multiplications (which is the case for almost all `Health-KEFIR` measures). Then the function f_i that expresses measure M_0 via M_i can be written as

$$M_0(S_0, D) = A(S, D) \times M_i(S, D) + B(S, D) \tag{20.8.5}$$

where $A(S,D)$ and $B(S,D)$ depend on the sector, the database instance, and other measures, but *not* on M_i. Now, the reference value of M_0 is

$$M_0(S_0, D_R) = A(S, D_R) \times M_i(S, D_R) + B(S, D_R) \tag{20.8.6}$$

Let $D_R^{(i)}$ denote the version of D_R where measure M_i in S has the value $M_i(S, D_C)$ (its value in D_C), and the rest of the measures are unchanged. The value of M_0 would be

$$M_0(S, D_R^{(i)}) = A(S, D_R) \times M_i(S, D_C) + B(S, D_R) \tag{20.8.7}$$

Subtracting, we get

$$impact(M_i, S, D_C, D_R) = A(S, D_R) \times (M_i(S, D_C) - M_i(S, D_R)) \tag{20.8.8}$$

which is the change in measure M_i in sector S, multiplied by the impact factor $A(S, D_R)$. Note that $B(S,D)$ – the contribution resulting from additive terms – drops away completely. This equation allows one to compute impact for measures related to M_0 by additions and multiplications simply by keeping track of the multiplicative factor $A(S, D_R)$.