Critical analysis reporting environment (CARE)

A. Parrish¹, D. Brown¹, R. Stricklin¹ & D. Turner²
¹CARE Research & Development Laboratory
²University Transportation Center for Alabama, The University of Alabama, USA

Abstract

The Critical Analysis Reporting Environment (CARE) software identifies traffic safety problems, evaluates countermeasures, finds hazardous sites, and provides investigative data. It is versatile and user friendly. Its data-mining feature allows traffic safety officials to use it even though they possess only a fundamental understanding of statistical theory, the software or the data set.

1. The user identifies a data subset of concern (alcohol use, speed, etc.) and creates a filter to define the corresponding subset of crash data.
2. The data mining module compares the filter subset against its data complements (or any other defined data subset) over all attributes.
3. The data mining module generates a comparison for all attributes in the database and arranges the output in the order of most significant attribute first. All statistically significant differences are indicated.

Since the output is listed in “worst first” order, the user can quickly identify the most productive work area. Using the data mining feature, safety professionals can then “drill down” into the data to find trends and answers to safety questions.

1 Introduction

The Critical Analysis Reporting Environment (CARE) was designed and developed at the University of Alabama for traffic accident problem identification and evaluation, with major sponsorship by the State of Alabama and the National Highway Transportation Safety Administration (NHTSA) [1]. CARE is nationally recognized, having won the NHTSA “Administrator’s
Award” for its innovative information mining capabilities. It has been implemented in Alabama, Michigan, North Carolina, Florida, Tennessee, Iowa and Delaware. At least three federal agencies use it, along with multiple cities. In Alabama, three state agencies use it in their safety programs, and over 200 individuals have been trained in its use.

The software provides analysts and decision-makers direct access to accident and incident information. It requires no formal training in computer hardware or software because CARE windows include options that provide a thorough guide to all desired output. Users can follow step-by-step directions to obtain information on the screen or at the printer [2].

2 CARE overview

The following provides a general description of CARE [3]:

- CARE exists in two platforms: desktop and Web. The desktop version operates on PC-compatible microcomputers under all recent versions of Windows (e.g., 95, 98, NT, 2000, ME, XP).

- Most CARE results come to the screen in a few seconds, providing feedback that stimulates the user to issue additional queries based on preliminary results. Once subsequent queries are refined enough to be printed, they can be saved to a file and edited with a word processor to produce the final report.

- CARE capabilities are extensible. It has satisfied well over 95 percent of the information requests for the safety databases to which it has been applied. For those few special studies that require more sophisticated statistical techniques, CARE can export files of user-defined data subsets as a front end for other statistical processors. Most exports are directly into Excel.

- CARE data comes from two sources: (1) downloaded from the central database, or (2) entered at the local level, ultimately providing an uploading capability.

- CARE provides several major advances that facilitate problem identification. Its rapid response to the user’s query encourages immediate modification of the query, giving the user the ability to hone in on exactly what he or she wanted. On the other hand, when an analyst does not know what is in the database, CARE gets the study started immediately through its information mining capability. This generates information by the comparison of subsets of data (e.g., weather-related vs. non-weather-related crashes), and graphically demonstrates potential areas for safety analyses.

2.1 CARE operational capabilities

This section outlines various CARE capabilities. They are presented in the general order in which users of CARE typically request them.

- Database/relation selection – This is the ability to select the database or the relation within a database from which information is desired.
Filter Selection – A filter is a specification that directs analyses at only a specific subset of the data. Certain filters are predefined because many of the subsets of interest can be anticipated in advance. For example, filters are generally predefined for traffic crashes caused by or related to: alcohol, bicycle, driver, emergency medical services (injury and fatal), fatalities, motorcycle, pedestrian, roadway defects, railroad, school bus, truck, vehicle defects, age, and political subdivisions (counties and cities), and any others for which analysis is expected.

Filter Combinations – Combinations of predefined filters can be created intuitively employing standard Boolean AND and OR operations. For example, the user might specify that only alcohol-related, motorcycle-fatality crashes will be analyzed, and create a filter from three existing filters. The analyst can form very sophisticated combinations using the simple user interface.

User-defined filters – In addition to using predefined filters, the user can create new filters. An intuitive interface walks the user through the process to select any combination of variables and values from the database to create a filter to define any subset. Examples could include age groups, blood alcohol levels, or driver visibility. Once a user-created filter is defined, it has the same status as any other predefined filter. It becomes part of the filter list (until deleted) and can be used in combination with other filters.

Frequency distributions – The simplest CARE analysis provides labelled frequency distributions for any or all variables for any subset. Variables (such as time of day, day of the week, weather, driver age, etc.) are chosen from a selection menu. The output includes a tabular frequency distribution and bar charts for visual analysis of the data.

Cross-tabulations – CARE produces fully-labelled cross-tabulations of sets of any two variables for any subset of the data.

Area Criticality Technique (ACT) – CARE produces a prioritized list (worst-first by rate), calculated from a defined demographic such as city population. Typically, this is composed of a list of cities (stratified by population groupings) that are prioritized according to crashes per city population for whatever subset of crashes is defined by the current filter.

Information Mining Performance Attainment Control Technique (IMPACT) This module performs true automated information discovery by systematically finding all over-representations between any two subsets. Graphical and tabular outputs are arranged in order of worst-first for each variable. This powerful tool finds and prioritizes over-representations without user knowledge of the attributes of the underlying database. As an example, in a recent aviation study, a comparison of weather-related accidents with non-weather-related accidents told the who, what, where (most over-represented locations), when, how and why so safety engineers could begin considering countermeasures for the most critical areas.
High-density locations – This module finds high accident roadway locations (intersections, non-mileposted segments, or mileposted segments) for any type of crash subset. The user can dynamically redefine the study limits (mileposted locations) to be sure that all relevant accidents are included for a location.

While some of the capabilities of CARE are quite sophisticated, they can all be obtained by merely selecting options from menus and following stepwise procedures. In addition, none of the procedures require more than 30 seconds of processing time, and most are returned virtually instantaneously. As an example, a complete IMPACT run over all variables for a database of a half-million records of 200 attributes takes less than 30 seconds on the typical desktop computer.

Although this paper deals mainly with traffic accident records, it is important to recognize that CARE is not restricted to this application. CARE can easily be adapted to new applications and to almost any type of dataset.

3 Case studies

Two case studies are used to illustrate the capabilities of CARE. The first illustrates the data mining “drill down” capability, and the second illustrates a practical application by a state agency to generate quick answers to a difficult, complex question.

3.1 Case study one – drill down feature

The State of Alabama recently adopted a “Graduated Driver License” (GDL) requirement that placed certain restrictions upon new (16-year-old) drivers to minimize their chances of being involved in a collision [4]. The logic was that the new drivers lacked experience, and that they should obtain it in several steps.

To guide GDL development and implementation, a study was conducted to compare 16-year-old drivers to their 17-20 year-old counterparts. Appropriate filters were designed and the IMPACT module was used to identify important crash factors. Time of crash was found to be the most significant variable when comparing the two groups. The tabular output of Figure 1 is in worst-first format, so the first line of the table was the most overrepresented (3:00-4:00 pm) and the next line (7:00-8:00 am) was the second-most overrepresented. In other words, the driving behaviour of 16-year olds was most different from the control group just before and just after school. The double-bar percentage chart clearly shows the same result, and is easier for a layperson to understand than the table.

The two right-most columns of the table provide important information. Asterisks in the “Over Represented” column designate statistically significant values. The “Max Gain” column gives the amount of overrepresentation (i.e., the number of crashes that could be prevented if 16-year-old driver performance could be modified to match 17-20 year-old driver performance).
In a similar manner, CARE found another high-priority variable to be the number of occupants in the causal vehicle. The 16-year-old drivers had a significantly higher number of occupants when compared to their 17-20 counterparts. This is very important since one of the factors in the GDL law was restriction of the number of vehicle occupants for 16-year-old drivers.

The CARE information mining capabilities are able to create information automatically, and to surface all relevant information from a database for subsequent analysis. Once an interesting item is found, instant insight can be obtained through a process called drilldown, which processes all of the variables of the database relative to the designated item. For example, one of the interesting variables in this study was over-representation of females in the 16-year-old category – interesting because it is quite rare that females are over-represented.

The drilldown feature was used to find the reason for the overrepresentation. “Female” was designated on the screen and the drilldown indicator was clicked. The results showed that females had particular problems in “Failure to Yield Right of Way.” A further drill down indicated that they were having the problems during left turns at low speeds, and that the problems were at stop signs and traffic signals (as opposed to yield signs). A final drill-down on stop signs indicated that they tended to be in residential areas as opposed to shopping or business districts. The series of continuous analysis steps resulted in a series of screens, each of which was saved for further analysis. This is illustrated in Figure 2.

Figure 1: Output screen for time of day.
This analysis was accomplished in a few minutes using CARE, and revealed a great deal of information about crash trends for 16-year-olds (related to their lack of experience). Any user can conduct this analysis, using intuition and his or her specific interests, and guided by CARE. The designers of CARE ensured that novice users can quickly learn how to operate it, providing on-screen icons to help them conduct such analyses.

3.2 Case study two – analysis of curvature/superelevation

A comprehensive research project was recently conducted for the Alabama Department of Transportation (ALDOT) to identify the relationships between a number of roadway characteristics and safety [5].

Traffic safety databases typically start with individual crash records compiled from law enforcement reports, and attempt to append useful data to them (such as highway geometric characteristics). Many safety studies have been conducted on this type of traffic safety database. However, there is an inherent deficiency in such studies because they cannot account for locations that have experienced no crashes. In other words, only sites that have experienced accidents are in the database.

To remedy this problem, a different file was created to model the entire ALDOT roadway system – two records for each 0.1-mile segment, one in each of its two nominal directions. This essentially created a roadway file to which
crashes could be written (called the roadway-crash file). Roadway characteristics data were also written to each segment from all available ALDOT files, including geometric information (number of lanes, lane width, shoulder width, curvature, etc.) pavement condition, pavement friction, average daily traffic, roadway identifier, and traffic crashes by severity.

In this file, the presence of crashes and the number of crashes by severity can be determined for any segment, enabling summaries for any subset of segments. Additionally, the roadway data codes were designed to facilitate visual identification of patterns associated with crashes. For example curvature codes were arranged from the sharpest inside curve to the sharpest outside curve, with tangents in the middle.

This comprehensive roadway-crash file allows detailed studies, which can be conducted quickly using CARE. One such study was conducted during the process of installing and testing the new file. It involved analysis of horizontal curves to identify locations where inferior superelevation might be contributing to traffic crashes. Prior to this project, identification of such locations would have required extensive manual comparisons of datasets to identify crucial relationships between curves and crashes, after removing the confounding effects of other data items.

The study was conducted quickly and automatically with CARE, using the roadway-crash file. A filter was created and locations were identified with sharp inner curves and positive superelevation, or sharp outer curves with negative superelevation. Care created the cross-tabulation in Table 1, which shows how curvature and superelevation are correlated over the entire state roadway system. This tabulation was produced in seconds, but would have required months of manual labor.

Ideally, the major portion of all segments would lie along a diagonal line from the upper-left to lower-right of Table 1, where inside curves have negative superelevation, and outside curves have positive superelevation. Segments with the opposite characteristics are seen in the upper-right and lower-left corners of the cross-tabulation. As “outliers,” they are candidates for further study.

Filters can be developed easily using CARE to isolate any or all combinations of curvature and superelevation. For example, any combination of the two variables that created the cross-tabulation in Table 1 can create a subset of the roadway system as the basis for more detailed analyses at individual sites.

A filter was set up to identify locations with the sharpest outer curves (smallest curve radii) and the steepest negative superelevation classifications. A threshold criterion of at least two crashes in 0.1 mile was used to identify the sites that most deserved further study. The resulting summary of roadway locations is shown in Figure 3. Specific locations for each of these roadways were determined by clicking on each of the roadways in the list. Then a filter was set up to consider the sharpest inner curves with the steepest positive superelevation classifications, and a similar site listing was produced.
Table 1: CARE Printout - Analysis of curvature-superelevation combinations.

<table>
<thead>
<tr>
<th>Severe Inside Curve</th>
<th>Heavy Inside Curve</th>
<th>Mod. Inside Curve</th>
<th>Slight Not Curve</th>
<th>Slight Outside Curve</th>
<th>Mod. Out. Curve</th>
<th>Heavy Outside Curve</th>
<th>Severe Outside Curve</th>
<th>Total %</th>
</tr>
</thead>
<tbody>
<tr>
<td>222</td>
<td>735</td>
<td>660</td>
<td>548</td>
<td>662</td>
<td>332</td>
<td>383</td>
<td>468</td>
<td>118</td>
</tr>
<tr>
<td>8.9%</td>
<td>6.3%</td>
<td>3.5%</td>
<td>2.0%</td>
<td>0.9%</td>
<td>1.2%</td>
<td>2.0%</td>
<td>3.9%</td>
<td>4.8%</td>
</tr>
<tr>
<td>436</td>
<td>1728</td>
<td>2359</td>
<td>2964</td>
<td>5400</td>
<td>2022</td>
<td>1563</td>
<td>1053</td>
<td>226</td>
</tr>
<tr>
<td>17.5%</td>
<td>14.8%</td>
<td>12.7%</td>
<td>11.0%</td>
<td>7.8%</td>
<td>7.6%</td>
<td>8.3%</td>
<td>8.8%</td>
<td>9.3%</td>
</tr>
<tr>
<td>616</td>
<td>3575</td>
<td>6623</td>
<td>10099</td>
<td>26700</td>
<td>8908</td>
<td>6104</td>
<td>3254</td>
<td>477</td>
</tr>
<tr>
<td>24.7%</td>
<td>30.7%</td>
<td>35.8%</td>
<td>37.7%</td>
<td>38.6</td>
<td>33.6%</td>
<td>32.7%</td>
<td>27.3%</td>
<td>19.6%</td>
</tr>
<tr>
<td>249</td>
<td>1958</td>
<td>4053</td>
<td>6210</td>
<td>18139</td>
<td>5961</td>
<td>3881</td>
<td>1955</td>
<td>200</td>
</tr>
<tr>
<td>10.0%</td>
<td>16.8%</td>
<td>21.9%</td>
<td>23.2%</td>
<td>26.3</td>
<td>22.5%</td>
<td>20.8%</td>
<td>16.4%</td>
<td>8.2%</td>
</tr>
<tr>
<td>134</td>
<td>1050</td>
<td>1997</td>
<td>3324</td>
<td>9220</td>
<td>3133</td>
<td>1969</td>
<td>1066</td>
<td>130</td>
</tr>
<tr>
<td>5.3%</td>
<td>9.0%</td>
<td>10.8%</td>
<td>12.4%</td>
<td>13.3</td>
<td>11.8%</td>
<td>10.5%</td>
<td>8.9%</td>
<td>5.3%</td>
</tr>
<tr>
<td>260</td>
<td>725</td>
<td>864</td>
<td>1182</td>
<td>3298</td>
<td>1517</td>
<td>1013</td>
<td>758</td>
<td>227</td>
</tr>
<tr>
<td>10.4%</td>
<td>6.2%</td>
<td>4.6%</td>
<td>4.4%</td>
<td>4.7%</td>
<td>5.7%</td>
<td>5.4%</td>
<td>6.5%</td>
<td>9.3%</td>
</tr>
<tr>
<td>91</td>
<td>334</td>
<td>477</td>
<td>744</td>
<td>1955</td>
<td>1207</td>
<td>835</td>
<td>518</td>
<td>120</td>
</tr>
<tr>
<td>3.6%</td>
<td>2.8%</td>
<td>2.5%</td>
<td>2.7%</td>
<td>2.8%</td>
<td>4.5%</td>
<td>4.5%</td>
<td>4.3%</td>
<td>3.9%</td>
</tr>
<tr>
<td>65</td>
<td>351</td>
<td>467</td>
<td>647</td>
<td>1719</td>
<td>1367</td>
<td>912</td>
<td>638</td>
<td>170</td>
</tr>
<tr>
<td>2.6%</td>
<td>3.0%</td>
<td>2.5%</td>
<td>2.4%</td>
<td>2.4%</td>
<td>5.1%</td>
<td>4.8%</td>
<td>5.3%</td>
<td>7.0%</td>
</tr>
<tr>
<td>206</td>
<td>632</td>
<td>677</td>
<td>804</td>
<td>1573</td>
<td>1551</td>
<td>1339</td>
<td>1179</td>
<td>331</td>
</tr>
<tr>
<td>8.2%</td>
<td>5.4%</td>
<td>3.6%</td>
<td>3.0%</td>
<td>2.2%</td>
<td>5.8%</td>
<td>7.1%</td>
<td>9.9%</td>
<td>13.6%</td>
</tr>
<tr>
<td>156</td>
<td>405</td>
<td>258</td>
<td>207</td>
<td>361</td>
<td>401</td>
<td>523</td>
<td>762</td>
<td>286</td>
</tr>
<tr>
<td>6.2%</td>
<td>3.4%</td>
<td>1.4%</td>
<td>0.7%</td>
<td>0.5%</td>
<td>1.5%</td>
<td>2.8%</td>
<td>6.4%</td>
<td>11.8%</td>
</tr>
<tr>
<td>Total</td>
<td>2490</td>
<td>11615</td>
<td>18484</td>
<td>26748</td>
<td>69156</td>
<td>26462</td>
<td>18654</td>
<td>11892</td>
</tr>
<tr>
<td>Percent</td>
<td>1.3%</td>
<td>6.1%</td>
<td>9.8%</td>
<td>14.2%</td>
<td>36.8%</td>
<td>14.0%</td>
<td>9.9%</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

Figure 3: Locations of sharpest outer curves and steepest negative superelevation.

The next step involved using the CARE high density feature (hotspot module) to identify the exact locations of these potentially dangerous curves. Then detailed analyses were conducted on accident types, rates and patterns using information readily available from CARE.

This analysis was conducted in hours (instead of months), and where justified, ALDOT identified roadway sites for modification. The capability to conduct comprehensive studies with such speed and accuracy allows optimum use of safety funding, and certainly reduces crashes and saves lives.
4 Comparison of CARE to existing systems

There are no software system competitors for some of the dedicated traffic safety/health features of CARE, such as the High-Density Location module. Nor is there a competitor data processor specifically configured (like CARE) to work on the types of problems found in the traffic safety/health field.

There are other software systems that can produce the sort of information provided by IMPACT. Examples include SAS (Statistical Analysis System) and SPSS (Statistical Processor for the Social Sciences). Both have comprehensive statistical capabilities; however, they lack the “self starter” and “self guided” nature of CARE. Considerable expertise and effort are required to apply them to a complex issue. The user must master the database (not easy in today’s world), examine and test the variables, review and compare outputs of work steps, then determine the next work step that is probably best. The IMPACT module does this instantly and automatically through data mining.

In effect, the ease of operation and the capabilities of CARE make it a superior data processor. Other systems cannot match capabilities like the following:

- Analyze all variables designated by the user, in one run.
- Display results in either natural or worst-first ordering; no other software allows the user to switch back and forth between ordering methods while searching for problem areas.
- Provide cost-benefit analysis by assigning valuation to all over-representations (max gain display).
- Order the output within each variable according to max gain, which highlights problems areas.
- Arrange all variables according to respective total max gains; so that the most important variables rise to the top.

5 Summary

This paper provided an overview of the Critical Analysis Reporting Environment (CARE) software, which identifies traffic safety problems, evaluates countermeasures, finds hazardous sites, and provides investigative data. The software is versatile and user friendly. Its data-mining feature allows traffic safety officials to use it even though they possess only a rudimentary understanding of statistics, the software, or the data set.

Two case studies illustrated the capabilities of CARE’s data mining module. The first featured the “drill down” capability, and the second was an application by a state agency to generate quick answers to a difficult, complex question.

Although this paper deals mainly with traffic accident records, it is important to recognize that CARE is not restricted to this application. CARE can easily be adapted to new applications and to almost any type of dataset.
Acknowledgements

The authors wish to acknowledge the support and encouragement of the National Highway Traffic Safety Administration, the Alabama Department of Transportation and the Alabama Department of Community and Economic Affairs in the design and development of CARE. The many individuals who participated in CARE’s development, implementation, and user training are warmly thanked. Finally, the authors exhort current users to continue to apply CARE to prevent traffic crashes and save lives.

References


