Data Mining Electronically Linked Part C-Part B Data Sets to Identify Usage Patterns and Predict Need

Dr. Deborah T. Carran, Ph.D.
Professor of Statistics in the Department of Teacher Development and Leadership in the School of Education at Johns Hopkins University

Dr. Linda Tsantis, Ed.D.
Professor and coordintor of the Early Childhood Special Education graduate program in the School of Education at Johns Hopkins University

Dr. John Castellani, Ph.D.
Professor in the Technology for Educators and Special Education Technology programs in the School of Education at Johns Hopkins University

Dr. Carol Ann Heath-Baglin, Ed.D.
Assistant State Superintendent Division of Special Education/Early Intervention Services Maryland State Department of Education
"Data does have a problem, in that it's only available about the past. (Clayton Christensen)

Introduction

During the past two decades as technology has advanced, the ability to link extant database records has become a reality. This ability has spurred the creation of large data warehouses at all levels of federal, state, and private organizations. The electronic linkage of state birth and school records have documented risk factors associated with child outcomes (Scott, Mason, & Chapman, 1999). Researchers have forged a new area in developmental epidemiology, which has provided excellent information for the investigation of risk factors at both the sample and population level of analysis. This is good for researchers, but are there other uses of these data by stakeholders?

Data mining is defined as “an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the patterns [models] to new subsets of data” (Statsoft, 2001, p. 1). Massive amounts of data sit in repositories that states do not know what to do with and have not been able to use. This paper introduces data mining as a tool for state level special education data (collected under federal mandate) to (a) identify and forecast the need for future special educational services and personnel, and (b) examine patterns in the data which may improve the efficacy of services for children with special needs.

History

In 1986, Congress enacted P. L. 99-457, adding Part H to the Education of the Handicapped Act (EHA, now the Individuals with Disabilities Education Act [IDEA]) “with the goal of encouraging development or expansion of statewide early
intervention services for children birth through 2 years with disabilities and their families” (U. S. Department of Education, 2000, p. II-1). In September of 1994, all states reported full implementation of Part H (renamed Part C in the IDEA amendments of 1990). The overall purposes of P. L. 99-457 were to insure coordination of early intervention services (EIS) for young children and their families, to provide access to a comprehensive interagency system of services, to maximize funding, to promote the development of innovative service models, to enhance training opportunities for parents and professionals, and to enact policies designed to be supportive of early intervention programs. Prior to P. L. 99-457, a variety of public and private agencies were providing a wide range of EIS through a diverse network and Congress recognized the need to coordinate services yet incorporate the preexisting state diversity (Cameto, et al 2000).

The Maryland Infants and Toddlers Program (MITP) is Maryland’s EIS which is implemented statewide, consistent with the requirements of Part C of IDEA. MITP is locally administered and implemented within the framework of the federal and state requirements. EIS include the variety of services, which have traditionally been available through special education, as well as more family-oriented support services. The State of Maryland provides EIS to infants and toddlers through funding and services available within the local departments of health and social services, and through LEAs.

In 1989 Maryland limited its definition of eligibility. To be considered eligible for EIS, an infant or toddler must (a) be experiencing at least a 25% delay in development in one or more areas, (b) manifest atypical development or behavior that is likely to result in subsequent delay, or (c) have a diagnosed physical or mental condition that has a high probability of resulting in delay (i.e. chromosomal abnormalities, genetic or congenital disorders, sensory impairment— including vision or hearing, inborn errors of metabolism, microcephaly, severe attachment disorders—including failure to thrive, seizure disorders, and fetal alcohol syndrome. It is evident that children receiving Part C services are at greater risk to experience later developmental delays which involve services delivered under Part B special education.
The numbers of referrals and number of children eligible for EIS under MITP have increased steadily over the years. In 2001, a total of 4,984 infants and toddlers and their families were receiving early intervention services with an individualized family service plan. During this same year, 112,400 students ages 3 to 21 received special education and related services (Part B) in Maryland. It is estimated that less than 10% of students receiving Part B services participated in preschool special education.

As the number of children receiving services in Part C and B have increased every year, so too have the need for personnel to deliver the services. Planning for these needed personnel is typically done through a deficit evaluation model, where administrators look at what was needed for the last year and put this need into next year's budget. Current (that year) ratio of personnel to children are calculated, everyone is shocked at how few service providers are available for all the children needing services, and further angered by hiring and budget freezes. This cycle repeats every year. If instead, models could be developed to forecast an estimated number of children who would likely need specific services (i.e. physical therapy) four years from today, provider hiring and budget planning could be made easier.

Administrative planning for staffing is important, but it is not a priority for the federal government. Services are mandated to help children and families. Under a model of preventive intervention, providing EIS may reduce the number of students with disabilities in need of costly Part B special education services (Ramey & Ramey, 1992). To determine efficacy, however, requires the linkage of records and longitudinal tracking with outcome comparison of students.

*Part C-Part B Data Linkages*

The National Monitoring and Promising Practices Web Site (http://interact.uoregon.edu/wrrc/monitor/profmonitoring.html) recently posted the results from their national survey Profiles of State Monitoring Systems-2001. Questionnaire data from 46 states found that only 5 (11%) states responded
affirmatively to the question “Our State monitoring system integrates Part B and Part C.” This is the current level of data linkages between programs. Over time both Part C and Part B may enroll the same students, but the programs do not integrate or exchange information. This is the major reason for the lack of broad information on efficacy of EI/S. These programs operate autonomously, exchanging information only when Part C children are referred for Part B continuation services.

Maryland, one of the 5 states reported above, is investigating the application of integrated Part B and C information. Several research questions have been formulated and are being addressed. These questions include:

- Do children who received Part C services in MD receive Part B services in MD?
- For children identified as receiving Part C-B services in MD, what changes occur over the years?
- How effective are Part C services on child outcome?
- Are parents satisfied with Part C and Part B services?
- What are the strengths/weaknesses in Part C-B outcomes, which may improve efficacy or delivery of special education services?

The first 4 questions listed above are primarily descriptive in nature. Careful data management and analysis will be required to investigate these outcomes, but they are easily accomplished with planning. The last question has different implications when considered with the services provided in both Part C and B. Specifically, the question may be rephrased as:
What Part C services are delivered to students who later receive specific Part B services? Are services or patterns of services effective?

Are there patterns in the data that may be used to plan for future provision of services for Part B special education?

The rephrased questions now imply forecasting future usage patterns based on historical patterns of Part C services delivered. With the linkage of the extant Part C with Part B data repositories, the pattern of Part C services could be the key to insuring that teachers and service providers are available to students when they enter Part B special education. Additionally, with further investigation providers could be more confident in the efficacy of prevention/intervention services.

If these questions could be answered, results would be remarkable, especially if linked to student outcomes. Prediction mandates statistical modeling and there are techniques of iterative statistical modeling which are data driven. One of the most effective applications comes from the emerging science and industry of knowledge discovery in databases, also known as data mining.

Data Mining

Data mining applies modern statistical and computational technologies to the problem of finding useful patterns hidden within large databases. To find these patterns some general steps to be followed may include (1) a query is directed about which variables predict/relate to an outcome, (2) models are built and validated that consist of independent variables and a dependent variable, and (3) the model with the least error is selected and interpreted. Data mining has proven to be a powerful tool capable of providing highly targeted information to support decision-making and forecasting for scientists, physicians, sociologists, the military and business. The predictive power of data mining comes from its unique design—it
combines techniques from machine learning, pattern recognition, and statistics to automatically extract concepts, and to determine interrelations and patterns of interest from large databases (Edelstein, 1997).

One important concept that sets data mining apart from other techniques is that it is not standard data analysis. Traditional theory driven data analysis in social sciences begins with the theory, generates research hypotheses, collects data to test the hypotheses, and interprets the results of the analyses as supporting/not supporting the theory. Data mining is a process that traditionalists may view as ‘fishing for results,’ but it is not. It is an iterative process where queries are formulated, models built from select data, tested (validated), then “revised until a meaningful predictive model evolves” (Edelstein, 2003, p.2). Data mining uses the data to drive the results and the results have been very fruitful.

Some recent examples of data mining have been the Human Genome Project database, which has analyzed gigabytes of data on the human genetic code. The Sky Object Catalog from a major astronomy sky survey that consists of billions of entries measured in terabytes. The NASA Earth Observation System which is projected to generate approximately 50 gigabytes of remotely sensed image data per hour. Needless to say, there is a need to intelligently identify and simultaneously analyze data for ‘nuggets’ of useful knowledge (Fayyad, Pitesky-Shapiro, Smyth, & Uthurusamy, 1996).

The most appealing aspect of data mining is what it produces—complex statistics transformed into usable visualization charts that convey large amounts of information to a user. The user may look at the structure of the visualization chart and formulate a logic or statistical model for monitoring or prediction.

Use of a model for purposes of monitoring will alert the user when patterns deviate from outcomes. For example, if your credit card number is used with a different purchase pattern, it may be ‘marked’ in the database. If it is used for merchandise you do not usually purchase, in areas of the country that you do not visit, and at times that you do not typically buy, this
purchase pattern will be marked by the data mining program as not typical/predicted/expected, and the credit card number made invalid.

For purposes of decisions, the models developed from the visualization charts may be employed to estimate outcomes modeled on the query variables. Once the structure of the visualization chart is determined, the problem query is posed. Data mining requires that queries be well formulated, not just fishing expeditions. The queries must specify the databases and variables of interest for the investigation. Then the data mining software may be used to analyze and form classification and regression trees, neural networks, rule induction, or nearest neighbor validation. Each of these techniques is a statistical methodology used to examine the relationships among variables. The technique chosen for this paper was Classification And Regression Trees (CART), and will be further explored in the following sections.

Data mining solves problems by analyzing extent data. It discovers meaningful patterns in data that may help explain something about the data. It has long been used for business, industry, and other groups but is only now being applied to education. Our goal in this paper was to use data mining to examine the existing relationship between Part C and Part B services for the purpose of predicting future usage patterns and improving services for children with special needs.

**Method**

**Participants**

All children who received Part C services in Maryland Infants and Toddlers Program (MITP) between the years 1995 and 2001 and who were receiving Part B services in Maryland for the 2001 academic school year and the 2002 school year were included in this study. The sample was limited due to lack of a common unique identifier across data sets. Part B required
social security number (SSN) for registration but Part C made SSN optional. Unfortunately, most jurisdictions did not record SSN for Part C registration. Between 1995 and 2001 in Maryland, there were 33,605 students served by Part C. Part B served a total of 112,400 children in 2001 (Part B_2001) and 113,199 served by Part B in 2002 (Part B_2002). Matching students in Part C and B_2001 using SSN, data of birth, and sex resulted in a matched sample of 3,060 and matching Part C and B_2002 identified 3,692 children (see Figure 1). There may be an overlap in children from the Part C database between the training and validation data sets, but they have different Part B outcome (dependent attributes). Two data sets were required for model validation. The 2001 data set was used for final models.
Figure 1. Schematic Merging of Part C and B Databases for Training and Validation Data Set in Data mining.
*C Children for each year includes data from C Biological Risks, C Environmental Risks, C Suspected Delays, and C Services

Data Mining

Data Preparation, Part C Predictor Attributes. Due to redundant data rows for multiple services in the Part C database, all Part C data files were ‘flattened’ by subject. Participants then linked data across years, 1995 – 2001. Part C data was often longitudinal, as children may remain eligible for services from birth through 3 years of age. Therefore, children may have received multiple services for up to 3 years. Data files merged into the Part C set included:

- Part C biological risk
- Part C environmental risk
- Part C Suspected delay
- Part C services (1992 – 2001)

See Table 1 for a listing of each data set. Part C service files were truncated to a complete listing of the first 4 services each year due to space limitations. All files were screened for field outliers/errors.

Table 1

Part C Risk Factors Services as Predictor Attributes and Part B Services as Dependent Attributes for Data Mining.
<table>
<thead>
<tr>
<th>Predictor Attributes</th>
<th>Predictor Attributes</th>
<th>Predictor Attributes</th>
<th>Predictor Attributes</th>
<th>Dependent Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>C Services</td>
<td>C Biological Risks</td>
<td>C Environmental Risks</td>
<td>C Suspected Delays</td>
<td>B Services</td>
</tr>
<tr>
<td>Assessment</td>
<td>APGAR &lt; 6 at 5 min</td>
<td>Disturbance in</td>
<td>Atypical development</td>
<td>Classroom Instruction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>relations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assistive Technology</td>
<td>Birthweight &lt; 1500 grams</td>
<td>Inadequate/unsafe conditions</td>
<td>Cognitive delay</td>
<td>Physical Education</td>
</tr>
<tr>
<td>Audiology</td>
<td>Congenital infection</td>
<td>Maternal age &lt; 15 years</td>
<td>High probability condition</td>
<td>Speech/Language</td>
</tr>
<tr>
<td>Case Management</td>
<td>Drug exposed/affected</td>
<td>Maternal mental retardation</td>
<td>Not specified</td>
<td>Travel Training</td>
</tr>
<tr>
<td>Family</td>
<td>Exposure to toxic substance</td>
<td>Maternal substance abuse</td>
<td>Physical delay</td>
<td>Career &amp; Tech Ed Program</td>
</tr>
<tr>
<td>Family Counseling/Training</td>
<td>Feeding dysfunction</td>
<td>Other</td>
<td>Psychosocial delay</td>
<td>Special Career &amp; Tech Ed</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------------------</td>
<td>-------</td>
<td>-------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Health</td>
<td>Gestational age &lt; 34 weeks</td>
<td>Prior involvement with PS</td>
<td>Self-help skills delay</td>
<td>Vocational Evaluation</td>
</tr>
<tr>
<td>Medical</td>
<td>Hearing impairment</td>
<td>Speech/Language delay</td>
<td>Special Education Program</td>
<td></td>
</tr>
<tr>
<td>Nursing</td>
<td>Metabolic disorder</td>
<td></td>
<td>Audiological Services</td>
<td></td>
</tr>
<tr>
<td>Nutrition</td>
<td>Neurological problem</td>
<td></td>
<td>Psychological Services</td>
<td></td>
</tr>
<tr>
<td>Occupational Therapy</td>
<td>Other</td>
<td></td>
<td>Occupational Therapy</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>Significant medical problem</td>
<td></td>
<td>Physical Therapy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Small for</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Therapy</td>
<td>gestational age</td>
<td></td>
<td>Recreation</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>----------------</td>
<td>---</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Psychological Services</td>
<td>Vision</td>
<td></td>
<td>Early Identification &amp; Assessment</td>
<td></td>
</tr>
<tr>
<td>Respite Care</td>
<td></td>
<td></td>
<td>Counseling Services</td>
<td></td>
</tr>
<tr>
<td>Social Work</td>
<td></td>
<td></td>
<td>Medical Services</td>
<td></td>
</tr>
<tr>
<td>Special Instruction</td>
<td></td>
<td></td>
<td>School Health Services</td>
<td></td>
</tr>
<tr>
<td>Special Day Care</td>
<td></td>
<td></td>
<td>Social Work Services</td>
<td></td>
</tr>
<tr>
<td>Speech/Language</td>
<td></td>
<td></td>
<td>Parent Counseling &amp; Training</td>
<td></td>
</tr>
<tr>
<td>Transportation</td>
<td></td>
<td></td>
<td>Other Therapies</td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td></td>
<td></td>
<td>Rehabilitative Counseling</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Orientation &amp; Mobility Training</td>
<td></td>
</tr>
</tbody>
</table>
Note: Children may have been eligible to receive multiple services across multiple years. Only the first 4 services listed for each year were used in datamining.

Data Preparation, Part B Dependent Attributes. Part B data were taken for the 2001 and 2002 academic school years. Data files for each year were merged into two Part B sets that included:

- Part B services 2001
- Part B services 2002

Data Management and Analysis. Part C and Part B data were electronically matched and linked on SSN and date of birth (see Figure 1). The individual Part Cs and Bs databases were merged to create two large data sets in MS Sequel Server (one for each year 2001 and 2002). Each resulting merged data file contained 464 variables. The two data sets were created based on electronic linkage for Part B_2001 data (N = 3060 children and 125,400 services) and Part B_2002 data (N = 3692 children and 117,246 services). One data set (Part B_2001) was used for model training and the other (Part B_2002) for validation. Part B_2001 data set was used for final data mining results.

Table 2

Training and Validation Services, Percent Outcomes for Samples and Final Models.
<table>
<thead>
<tr>
<th>B Service</th>
<th>Training Sample</th>
<th>Model Nodes</th>
<th>Model Training Results</th>
<th>Model Validation Results</th>
<th>Difference&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% No</td>
<td>% Yes</td>
<td>% No</td>
<td>% Yes</td>
<td>% No</td>
</tr>
<tr>
<td>Classroom Instruction</td>
<td>30</td>
<td>70</td>
<td>4</td>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>Occupational Therapy</td>
<td>67</td>
<td>33</td>
<td>5</td>
<td>67</td>
<td>33</td>
</tr>
<tr>
<td>Physical Therapy</td>
<td>85</td>
<td>15</td>
<td>3</td>
<td>85</td>
<td>15</td>
</tr>
</tbody>
</table>

<sup>a</sup>Difference between %correct classification rate of training and validation models.

Once data sets were merged, selection was made of outcomes or dependent attributes for data mining. Since all 23 Part B services could be dependent attributes for data mining, decisions were made to narrow the field for this initial exercise. All services had 2 levels: Yes (service delivered) or No (service not delivered). The first decision was to select dependent attributes from the linked data sets with a minimal distribution of 15% in one level. The second decision was to select dependent
attributes that could be impacted by personnel usage patterns (e.g. therapists, teachers, etc...). Part B services selected from 2001 for data mining were Physical Therapy ($N = 445, 15\%$ Yes), Occupational Therapy ($N = 1025, 33\%$ Yes), and Classroom Instruction ($N = 2153, 70\%$ Yes) (see Table 2). These three dependent attributes present a range of frequencies while meeting the initial decision criteria.

Data Mining Procedure, CART construction. The term Classification Tree is used when the dependent attribute is categorical and the goal is to classify cases into categories, with a minimal error rate. In this analysis, the term ‘tree’ is more than a metaphor. Results are presented in tree-like figures (see Figures 2 and 3) with branches representing the splitting of cases based on values of predictor attributes. The branches end in Nodes where either another predictor attribute is added to the tree, or the branch terminates in a Terminal Node with final classification result of the dependent attribute. This discussion will focus only on binary decisions trees.
Figure 2. Part B Physical Therapy Services (Dependent Attribute) Classification Tree with Part C Services as Predictor Attributes
Figure 3. Part B Classroom Instruction (Dependent Attribute) Classification Tree with Part C Services as Predictor Attributes

Once the data were in Sequel Server and the query defined C Service predictor attributes and the B Service dependent attributes, the data mining process was started. There were three steps in model screening prior to data mining: (a) formulation of the tree through training and validation, (b) pruning of the tree to enhance classification rate, and (c) visualization of lift charts to determine optimal tree. Training, validation, pruning, and visualization of lift charts will be explained in this section. CART results will be presented and interpreted in the following section.
Training & Validation. The use of one data set to build (train) a classification tree (model) and using another data set to validate the model is one method of model formulation. In this way, the tree is built (trained) in a stepwise fashion (i.e. statistically using a split selection method, not theoretically) with only statistically significant variables entered into the classification tree. The model is then validated with the other data set to determine if the untrained data ‘fit’ the model. A good ‘fit’ is one with little difference in classification percentage between training and validation.

For purposes of this study, the training data consisted of all C services from 1992 to 2001 matched to Part B_2001 (N = 3060). The validation data consisted of all C services from 1992 to 2001 matched to Part B_2002 (N = 3692).

Tree Pruning. During the training phase, a large classification tree is generated with many branches. This results in severe overfitting, with the consequence of capturing noise in the model. Restricting the model through pruning forces the analyst to review the tree and its node components and make decisions about combining or eliminating one or more nodes. The goal, as always, is to maintain the highest classification proportion with the least number of variables.

For this study, initial pruning was based on apriori decisions of smallest acceptable cell sizes and misclassification rates. Terminal cells with less than 200 subjects (6% of population) were pruned and regrouped with other larger nodes.

Lift Charts. Once the classification tree has been selected and pruned, additional analyses may be done to support the robustness of the tree. The lift curve in data mining is a visualization technique used to gauge performance. With a validation data set that is used to classify cases into one of two outcomes, a probability may also be calculated for each case as to its probability of belonging to each dependent attribute class. The higher the curve deviates from the linear 90 degree diagonal on the face of the chart, the better the tree in predicting outcome. Lift charts were used to judge the efficacy of each model to enhance classification rates.
Statistical Analysis
The final statistical analysis in Data Mining is the production of CART trees. The trees produced by the software follow close scrutiny by the analyst in the steps outlined above. Outcome in percentage of correct classification will be interpreted. The Classification And Regression Tree (CART) software program (Sal, 2004) was the data mining tool.

Results

Using the three dependent attributes of Part B Physical Therapy, Part B Occupational Therapy, and Part B Classroom Instruction, three complete data mining analyses were conducted. The results of each step in the process will be described.

Model Screening. Table 2 presents the final training and validation models for the dependent attributes. These models had previously been pruned and analyzed with lift charts to arrive at this end stage of screening. As defined earlier, a good ‘fit’ is one with little difference in classification percentage between training and validation. From Table 2 it was determined that the model for Classroom Instruction presents the best fit between training and validation (2% difference), Physical Therapy had a 7% difference, and Occupational Therapy had a difference of 11%. Model selection was further supported through inspection of the lift charts. Each lift chart indicates model improvement during training and validation beyond no model. Improvement was demonstrated by deviation from the 90 degree diagonal. Figure 4 the lift chart for Classroom Instruction demonstrated an increase in correct classification rate for the model compared to no model. The lift chart, in addition to the strong correct classification rate between training and validation models, supported final model selection for Classroom Instruction. As the researchers reviewed lift charts and discrepancy between training and validation data, this final model for Classroom Instruction was decided to be the best algorithm and the combination of predictor variables were logical.

The lift chart for Physical Therapy (see Figure 5) presented a visually convincing model. While there was a 7% discrepancy
between training and validation, the model dramatically increased the correct classification for Physical Therapy compared to no model (as determined by the dramatic deviation from 90 degrees with the training and validation models). This model was determined by the researchers to be reliable.

The final Part B service predicted was Occupational Therapy. As mentioned above, there was an 11% discrepancy between training and validation, yet the lift chart (see Figure 6) looked slightly better than Classroom Instruction. However, with a discrepancy of greater than 10% between training and validation this model for Occupational Therapy the a priori decision rule made this model not reliable.

Based on these results, CART results will be presented and interpreted for Part B Physical Therapy (see Figure 2) and Part B Classroom Instruction (see Figure 3).

**CART Results, Physical Therapy.** The dependent attribute is Part B Physical Therapy Services (categorical, yes or no) and the predictor attributes entered into the training model were all Part C variables listed above. At the top of Figure 2 is the Root Node (Node1) where the first split occurs. The software program reviewed all predictor attributes and the variable selected with the best resulting classification on the dependent attribute (minimum misclassification) was selected as the Root Node, Part C Physical Therapy Services 1 year. Children who received at least 1 year of Part C Physical Therapy services (Yes) split into Node 2 (n = 1008, 33%) and children who did not receive at least 1 year of Part C Physical Therapy services (No) split into Terminal Node 1 (n = 2052, 67%). These 2 groups of children present different outcomes and different patterns of services. The Root Node variable, Part C Physical Therapy, proved to be a very strong discriminator of the outcome, Part B Physical Therapy services.

The results of Terminal Node 1 were that 97% (n = 1994) of children who did not receive Part C Physical Therapy (n = 2052) also did not receive Part B Physical Therapy. This split ended in Terminal Node 1, since no additional variables could improve
upon the 97% correct classification rate.

For children who received Part C Physical Therapy \((n = 1008)\), additional variables were determined to be significant discriminators. At the next step (Node 2), remaining predictor attributes were again reviewed by the program and the predictor variable with the best resulting classification, Part C Occupational Therapy 1 year, was selected. As shown under Terminal Node 2, nearly 75% \((n = 272)\) of children who received Part C Physical Therapy \((Yes \ C_{\ PT})\) but did not receive Part C Occupational Therapy \((No \ C_{\ OT})\) did not later receive Part B Physical Therapy; 25% of children with these same service patterns \((Yes \ C_{\ PT} \ and \ No \ C_{\ OT})\) later received Part B Physical Therapy.

Terminal Node 3 children who received both Part C Physical Therapy \((Yes \ C_{\ PT})\) and Part C Occupational Therapy \((Yes \ C_{\ OT})\), 45.8% later received Part B Physical Therapy; however, 54.2% of children who received the same pattern of services \((Yes \ C_{\ PT} \ and \ Yes \ C_{\ OT})\) did not require Part B Physical Therapy services.

Based on this tree, the pattern of services which best predict students receiving later Part B Physical Therapy services was the path to Terminal Node 3: If students received at least 1 year of Part C Physical Therapy \((Yes \ C_{\ PT})\) AND Part C Occupational Therapy \((Yes \ C_{\ OT})\), 46% later received Part B Physical Therapy services.

The best pattern of not receiving Part B Physical Therapy services was the path to Terminal Node 1: If students did not receive Part C Physical Therapy \((No \ C_{\ PT})\), 97% did not receive Part B Physical Therapy services.

*CART Results, Classroom Instruction.* The dependent attribute is Part B Classroom Instruction Services (categorical, yes or no) and the predictor attributes were Part C variables listed above. At the Root Node (Node1) the variable with the best resulting classification on the dependent attribute was Part C Special Instruction Services 1 year. Children who received at least 1 year of Part C Special Instruction services \((Yes \ C_{\ SI}, \ n = 2021, 66\%)\) split into Node 3 and children who did not receive
at least 1 year of Part C Special Instruction services (\(\text{No C\_SI}, n = 1039, 34\%\)) split into Node 2.

At Node 2, 2 years of Part C Speech Language services was added. Terminal Node 1 is not very informative since the lack of Part C services (\(\text{No C\_SI}\) and \(\text{No 2C\_SL}\)) resulted in a large proportion (61.5\%) of children later receiving part B Classroom Instruction. For those children landing in Terminal Node 2 who did not receive Part C Special Instruction (\(\text{No C\_SI}\)) AND did receive 2 years Part C Speech Language (\(\text{Yes 2C\_SL}\)), 67.1\% later did not receive Part B Classroom Instruction services while 32.9\% of children with the same pattern of services received Part B Classroom Instruction.

Node 3 adds another level, Part C Physical Therapy services, for children who received Part C Special Instruction. Nearly 90\% of Children who received Part C Special Instruction (\(\text{Yes C\_SI}\)) AND Part C Physical Therapy (\(\text{Yes C\_PT}\)) later received Part B Classroom Instruction. This was the highest level of classification for children receiving these services and also identified 34\% (\(n = 735\)) of the total Part B Classroom Instruction sample.

The final variable added at Node 4 does not add much information to discriminate beyond that of the Root Node. Nearly 80\% of children who received Part C Special Instruction (\(\text{Yes C\_SI}\)) AND not Part C Physical Therapy (\(\text{No C\_PT}\)) AND not Part C Speech Language 2 years (\(\text{No 2C\_SL}\)), later received Part B Classroom Instruction.

Based on this tree, the best pattern for students receiving Classroom Instruction services is Terminal Node 5: If students received Part C Special Instruction (\(\text{Yes C\_SI}\)) AND Part C Physical Therapy (\(\text{Yes C\_PT}\)), 90\% received Part B Classroom Instruction services. The best pattern of not receiving Part B Classroom Instruction services is Terminal Node 2: If students did not receive Part C Special Instruction (\(\text{No C\_SI}\)) AND received 2 years Part C Speech Language (\(\text{Yes 2C\_SL}\)), 67\% did not receive Part B Classroom Instruction services. Terminal Node 1 also presents a large proportion of children who received neither Part C Special Instruction nor Part C Speech Language, yet later received Part B Classroom Instruction (61\%). This is an absence of service patterns, but not likely an absence of services.
Once a tree structure is identified for an outcome, a rule-based/logical system or statistical model may be used for prediction. The difference between the logical system and model depends on the resulting tree structure. A simple tree such as Figures 1 and 2 with few variables is one that strongly discriminates. This type of tree is best used as a logical rule-based system where a proportional number of outcomes are predicted at every split. A tree that is more complex with several variables may best be used as a model for prediction and classification with regression analyses. Implications are discussed below.

**Discussion**

Data mining was useful in answering some questions related to the linkage of Part C and Part B services and will clearly be of great assistance in the planning for Part B service and personnel needs based on Part C service usage patterns. Use of information from Figures 2 and 3 will be discussed. Not as clearly addressed were the identification of patterns in the data related to the efficacy of EIS. However, many implications will be discussed for consideration.

It is clear from Figure 1 that the majority of children who did not receive Part C Physical Therapy (67% of the total sample) did not later receive Part B Physical Therapy (Terminal Node 1, 97% of children not receiving C_PT). This is logical. If they do not need Physical Therapy early in their life, they are not likely to need it later. On the other side of this tree were children who did receive Part C Physical Therapy (33% of the sample). The Node 2 split Occupational Therapy was a good discriminator. For children who did not receive Part C Occupational Therapy (Terminal Node 2), only 25% later received Part B Physical Therapy. Children who received Part C Physical Therapy and Part C Occupational Therapy (Terminal Node 3), 46% later received Part B Physical Therapy. A total of 445 children later received Part B Physical Therapy in this sample. The reason this is a good model is that Terminal Nodes 2 and 3 (see Figure 2) together identify the majority (87%, n = 387).

This logic model may be used to forecast future need for physical therapists by estimating the proportion of children who...
may later receive Part B Physical Therapy. Specifically, determine the number of children exiting Part C who received Yes C_PT AND Yes C_OT (Terminal Node 3) and multiply this value by .458 and the number of children exiting Part C who received Yes C_PT AND No C_OT (Terminal Node 2) and multiply this value by .253. Add these two values together to obtain the estimated number of students who may later need Part B Physical Therapy. Based on this value of future need and estimated exit numbers from Part B Physical Therapy services, the number of children who will have a higher probability of needing Part B physical therapists may be estimated. Using these service patterns from Part C in addition to annual employment statistics could indicate trends in needs.

Building a similar logic model from Figure 2, patterns of services that discriminate Part B Classroom Instruction outcomes may be calculated. Following patterns of services to terminal nodes, outcomes with proportional values may be used to forecast future occurrences of need for Part B Classroom Instruction. Using logic models built on prior service patterns and employment statistics, estimates for service personnel are not just a shot in the dark. Rather they are databased decisions.

What is not evident in this examination of service delivery patterns is the extent of EIS efficacy or patterns of services that promote EIS efficacy. For example in Figure 3 at Terminal Node 1, 61.5% of children with the service pattern No C_SI AND No 2 C_SL later received Part B Classroom Instruction. Moving directly to the right in Figure 3 to Terminal Node 2, a similar proportion of children who received the 2 years of Part C Speech Language (No C_SI AND Yes 2 C_SL) later did not receive Part B Classroom Instruction. Did the 2 years of Part C Speech Language services prevent the need for Part B Classroom Instruction? Data mining at this general level will not answer this question.

This paper does not attempt to individually investigate every child’s service delivery pattern. This is a general statistical model, which bases results on overall total classification rate. However, this model does bring this question to the forefront. Based on data provided by patterns of services linked across years, questions like this may be investigated. This method may help to identify services or other variables that will eventually promote some level of prevention.
Limitations

A caveat of this study that the reader must keep in mind is that all of these children received both Part C and Part B services. There was no primary prevention model that was examined. There was no ‘cure’ of a condition which made these children eligible for Part C services. But we do want to examine patterns of services that may be important in secondary or tertiary prevention of need for specific future services.

A second limitation was that only matched Part C and B children were used for model training and validation. If a family or jurisdiction chose not to include a child’s social security number, that child could not be included in the study. This limitation impacted the number and types of services available in the training and validation data sets. For example, in 2001 only 5% of children received Part B PT services, but our B_2001 matched data set had 15%.

A third limitation was that the sample was biased by Part B participation during years 2001 and 2002. Records were not available to researchers for participants who had exited from the Part B program into regular education. The students who exited may have benefited the most from Part C services. It was not possible to determine from these data what proportion of Part C children received services that permitted them to begin school in a typical classroom.

The primary limitation of this study was the number of subjects. We simply did not have enough data to identify patterns for all outcomes we were interested in exploring. Three thousand may sound like a large number, but in reality we needed 5 - 10 times that number for sufficient cell sizes. Consequently, Part B outcomes which had a low proportionate distribution like Part B Psychological Services and Speech Language were inconclusive.

Several questions emerged with the data mining results that frame future research questions and areas for investigation:
Do we have a structural view of the model for some of these services? Yes, for some. While classification rates were not perfect for this pilot study, more parsimonious models using all the data were informative. For example Part B Classroom Instruction was best discriminated by Part C Period of Special Instruction service Year 1. Similarly, Part B Speech Language was first classified by Part C Location Speech Language services Year 1 had good results, but the distribution (80% received these C services) raised some critical policy questions.

What was the role of some services (e.g. case management)? A closer look should be taken at some of the services do determine if they evidence a magnifier effect or a mediation effect. Not everyone received this service. Does presence of this service enhance other services? Why don’t all subjects receive case management? Service specific questions need to be addressed when the numbers of linked services in the data warehouses reach sufficient size.

What is the role of other services, like Speech Language? Service data indicated that nearly everyone enrolled in Part B and Part C received Speech Language. This is understandable given the reluctance of school districts to label young students, but do more than 90% of students receiving Part C services need Speech Language? This is a call for closer monitoring and the longitudinal outcomes, which are also being investigated with other studies. This looks like a starting point for policy development.

Can a better data set be built? A web based format for Part C data entry is currently being implemented. Maryland State Department of Education/Department of Special Education has taken the initiative to create a registry for students, which will be used for both the Part C and Part B systems. This has been accomplished. With the new data repository design, MSDE will require all children screened to have one unique identification number for use by both Part C and Part B. While these programs have been isolated and separate since their inceptions, merging the data sets was extremely difficult and some data were lost in the process. The result has been a big step forward and will facilitate linking and tracking children for specific outcomes analyses.

How can techniques such as datamining and other knowledge discovery techniques be applied to student achievement in special education? Tsantis and Castellani (2001) have looked at this question and made a series of recommendations for
school districts and states. They have concluded that it will be necessary to restructure the way data and resources are documented in schools at the service provider, teacher and student level. These newly designed information and data repositories will be managed by a highly trained Knowledge Specialists. Only by properly managing the huge amounts of data sitting in repositories (and being appended daily) may relevant and useful individual information be available. This will provide special educators with powerful new decision making tools that will be needed to face complex teaching and learning challenges.

- Is monitoring possible with a new data set design and datamining? We would like to think that this would be a natural development. We are hopeful that the states would want this monitoring process to be synchronous with the data collection itself. Change is not helpful a decade later and does not foster improved outcomes for special needs students. Monitoring is most effective if it is current and based on known outcomes. It makes sense that the states would want to monitor the federally mandated requirements, which have been forced upon them. At least for a while. Data could be collected which might support or refute the necessity of such requirements. All types of policy issues could be evaluated and addressed with data based conclusions using this technique of data mining.

In summary, using datamining as an analytic tool on electronically linked data sets was very informative. It has helped to focus our research questions, identify areas in need of additional investigation, and guide the new Maryland state Part C web based data entry project. With new data that will be added this year and with the historical data, which has just been identified, it is hoped that the data set will include 8,000 children; not the 30,000 we estimate are needed for reliable and accurate model building, but it will be a good starting point for this project and for the children of Maryland.

References


U.S. Department of Education (2000). *Twenty-second annual report to Congress on the implementation of the Individuals*
with Disabilities Education Act, Washington, D.C.
Figure 4. Validation and Training Lift Chart for Classroom Instruction
Figure 5. Validation and Training Lift Chart for Physical Therapy
Figure 6. Validation and Training Lift Chart for Occupational Therapy