A Risk Focused Model for Improving Information Quality: Lessons from the Science of Epidemiology

Lane A. Ongstad
Colonel, USAF
Director, Patient Informatics Division
Office of the Surgeon General
Tel: 210.536.3982
Fax: 210.536.5167
E-mail: Ongstad_L@msa01.brooks.af.mil

A Risk Focused Model for Improving Information Quality: Lessons from the Science of Epidemiology

ABSTRACT

As organizations move into the information age we are increasingly aware that competitive advantage accrues to enterprises that can manage the quality of their critical data most efficiently, effectively, and economically. Unfortunately, to date most organizations have taken a distinctly "after the fact" approach to dealing with data quality – that is to say, they wait for operational problems to appear before refining processes to correct the problems. Facts of life such as the multiplicity of disparate legacy systems, existence of stovepipe systems, movement to client server networks, increases in data warehouses, globalization of business, and the exploding base of IS users within the organization have made many of these problems universal. The "fix what's broken" approach is understandable and to some degree unavoidable given these trends in the information systems world. Not withstanding these facts, competitive advantage will go to organizations that are able to produce quality information on the "first pass". The literature regarding quality management in manufacturing processes has long recognized the economic advantages of defect prevention. In the medical field the science of epidemiology has historically taken a parallel approach to disease prevention. i.e., it seeks to prevent expensive treatment for disease by understanding the underlying processes and precursers and intervening in that disease process so as to prevent the illness from ever occurring. The body of evidence supporting the techniques of epidemiological assessment and intervention are long standing and well understood. Furthermore, the process has been extended beyond the medical field for use in designing preventive strategies for a variety of societal problems. This paper applies that same approach to the management of information quality. The paper presents a theoretical basis for identifying potential data "pathology", it also presents a practical, structured approach for applying the theory in an operational environment. A currently ongoing case study of managing the biometric information quality within the office of the Air Force Surgeon General is presented to illustrate the use of this approach.

INTRODUCTION

In many ways the development of information management has paralleled the development of modern medicine in the later twentieth century. Both have seen an explosion in the growth of available

technologies and practices that defy the ability of either the producer or the consumer to fully comprehend, let alone control. The general consensus of popular opinion is that both have radically changed how we view our world and how we live our lives. And while no one seems to doubt the basic benefit of these technologies, there is tremendous concern about our ability to understand and manage that which we have created. In the field of medicine there is also great concern with the exploding costs of these technologies, our inability to correlate these costs to "quality outcomes," and the very real question of "what is quality?" Again, with reference to medical care, we have seen that most of these expensive technologies have been focused on resolving or ameliorating the effects of a disease process after it has advanced to the stage where it becomes disruptive to the patient. (Witness the expense involved in organ transplantation, etc.) Whether because we are entranced with the mysteries of exotic technology, more strongly motivated by current then possible catastrophes, or reluctant to spend resources on potential problems that do not exist in the present, comparatively few resources have been devoted to the prevention of disease vice the treatment. Whatever the reason, it is the economics of the marketplace that are beginning to reverse this trend. As medical care in America becomes more and more subject to market forces, medical care organizations are increasingly looking for means of controlling the costs of care and they are increasingly looking to the field of epidemiology for ideas.

Very briefly stated, the science of epidemiology seeks to understand disease processes. Using a structured, scientific approach, it identifies disease causes, precursors, and vectors. Armed with this knowledge it then seeks to intervene in the process so as to prevent the disease from occurring. Ideally interventions take place as far "upstream" from the potential onset of the disease as is possible. The premise is that the nearer the source an intervention takes place the more effective it is, the cheaper the cost, and the greater the potential savings. Is the approach effective? Without providing in depth research analysis, the point is well illustrated by the polio epidemic of the 1950s. There is no doubt that the development of an inexpensive vaccine, administered to the vulnerable population well before the onset of symptoms, successfully prevented tremendous future expenses involved with treatment. Moreover, aside from direct savings associated with avoiding treatment, it also produced additional indirect cost avoidance associated with the continued work productivity of those individuals who were spared the agony of contracting a debilitating disease. Moving further "upstream" the even more inexpensive cost of mosquito control undoubtedly had a tremendous cost benefit payoff over treating millions of cases of yellow fever.

Why does epidemiology succeed in disease prevention? Largely, because it takes a broader view of the disease process. It does not view the symptoms of disease in isolation nor does it seek to deal directly with the symptoms. Rather, it considers the larger system in which the individual (the person

who actually may contract the disease) exists and how the various aspects of their environment interact to produce disease. The approach also concentrates on the identification of predisposing risk factors in the environment and the mitigation of those risk factors as a means of avoiding undesirable outcomes (disease or illness). The results of this approach are a well documented, quantified, positive impact on both sides of the medical care quality equation: improved health (greater longevity, lower morbidity, etc.) at lower cost.

Recently these same systems oriented epidemiological principles and methodologies have been exported out of the medical field and used to design interventions to prevent a variety of community and social ills. Examples include firearm deaths[i], smoking[ii], school-related injuries[iii], drowning[iv], and alcohol related violence[v].

Our work in managing the quality of biometric data produced by medical systems led us to become familiar with the principles of preventive epidemiology. Further study, research, and experimentation led to the adaptation of these epidemiological methodologies and principles as a powerful tool for use in addressing data quality problems. The remainder of this paper presents the results of this work. It begins with a basic epidemiology model focusing on the ranges of prevention as applied to data and explores the competitive "leverage" that prevention provides. The paper then presents a model for identifying, assessing, and prioritizing specific data quality problems within an organization and demonstrates use of the model by presenting an actual case study using the tool within a medical organization. Finally, conclusions are drawn regarding the strengths and weaknesses, extent and limitations of the model as revealed in actual practice.

RANGES OF INTERVENTION

Before presenting a specific tool it is first necessary to gain a conceptual understanding of how and where risk identification and problem prevention / correction occurs within a system. A useful conceptual model is that produced by the Institute of Medicine entitled the "Range of Interventions" and shown in figure 1. Designed to explain the spectrum of possible interventions in the treatment of human disease, it is also directly applicable to the spectrum of risk factor identification and problem prevention in the treatment of data quality. As mentioned previously, the basic principle is that problem prevention is generally more inexpensive and more effective from a total systems standpoint than is treatment after the fact. Critical to understanding the applicability of this model to data quality management is a complete understanding of the model and its component fields. Although the field names are identified with medical terms that may be unfamiliar to the reader I will demonstrate their relevance to our data

quality field and translate them into understandable examples. Although the analogy is not perfect in all respects, it does correlate highly with basic DQ principles.

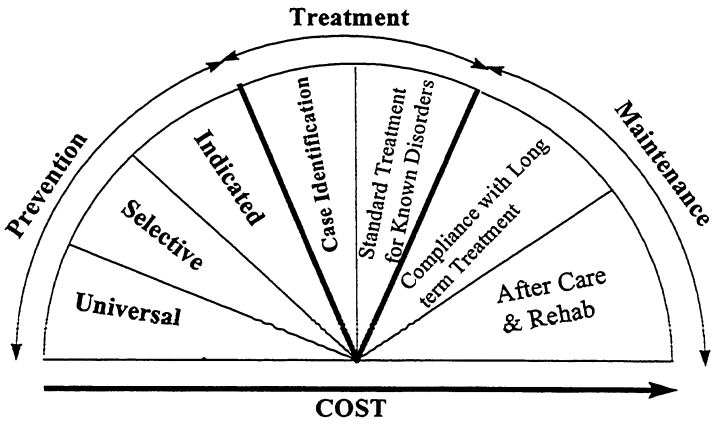

Figure 1[vi] IOM Range of Interventions

Overview: The model displays all possible interventions across the disease spectrum. Interventions range across three primary categories: prevention, treatment, and maintenance, with each category being further subdivided. These three categories of intervention encompass actions taken to forestall adverse outcomes (prevention), actions taken to deal with existent problems (treatment), and actions taken to prevent reoccurrence of past problems (maintenance). These can easily be extended to encompass actions (interventions) needed to deal with data quality problems as well. The cost arrow at the bottom of the model (not a part of the IOM model, added by the author) illustrates that the cost to the organization of correcting outcome (DQ) problems increases as interventions move across the spectrum. Unquestionably a problem avoided (prevention) is less expensive than correcting a problem that has occurred (treatment). (While this is obviously true, it is equally obvious to all those involved in operational DQ that cost justifying the resources necessary to prevent problems is also problematic). In an ideal world we would deal with all problems (medical or data quality) through prevention. In actuality, interventions to attain quality outcomes will range across the entire spectrum. The more practical goal then is to push as many problems as we can to the left (into the prevention range) through a proactive process that attempts to understand the processes at work, assess the risk factors present in the environment, and create "interventions" to decrease risk and therefor the potential for adverse results.

Prevention spectrum: Preventive actions are those interventions taken prior to the appearance of a problem. They specifically address issues and needs that, if not addressed, will result in problems (poor DQ). Preventive efforts also operate across a spectrum. In this case the spectrum ranges from general to specific.

Universal interventions are those actions which are most fundamental to disease prevention. Generally speaking they are the furthest "upstream" efforts, are the cheapest to implement on a per unit basis, and are the most effective in achieving prevention. Medical examples include inoculation, sanitation, and other public health efforts that are universally applied across an entire population in order to provide protection against disease. From an information systems perspective universal interventions include such basic measures as clearly stated business needs for data, clear policy on data capture, explicit and quantified standards for measuring DQ (a subject for another paper), good metadata, data metrics, benchmarks, active use of data, and ongoing executive involvement. Universal interventions are not appropriate for addressing all DQ problems but they can be viewed as a fundamental requirement for any data system – that is they are a necessary but not sufficient condition for DQ. Universal interventions are the foundation upon which the structure of an information quality program is constructed.

Selective interventions are those actions which are targeted at a specific sector of a population that is at high risk for contracting disease. These individuals occur in distinct groups and possess distinct characteristics that allow them to be readily identified and targeted. Since selective interventions tend to be more expensive than universal interventions it is neither economical nor effective to provide these interventions across the system. Medical examples include breast cancer screening for women within specific age groups, sickle cell screening for persons of African descent, and ophthalmic examinations for diabetics. Information systems examples may include targeted training, incentive systems, business practice improvements or particular systems interfaces prepared specifically for different organizational levels (headquarters, divisions, production units) or occupational specialties within the organization (data inputters, data users, analysts, etc.). Selective interventions address the unique needs of specific groups and sub-groups within the organization. They mitigate risk by focusing on the specific tasks and functions performed by the sub-group, anticipating common "failure points", and implementing protective measures. These protective measures encompass training, tools, processes, procedures, and policy support as needed to ensure that tasks are accomplished correctly and organizational expectations are met.

Indicated interventions are the most advanced, most expensive, and most targeted form of prevention. They are also the least understood and least deployed interventions in both the medical and information fields. Theoretically, they advance the science of prevention from groups to individuals

and attempt to detect individual risk factors and apply individualized preventive measures. Indicated medical interventions would include genetic screening and genetic manipulation, intrauterine surgery, etc. -- all very expensive on a per case basis but all avoiding huge "downstream" treatment and maintenance costs. Information systems examples would include artificial intelligence systems, adaptive technologies, fuzzy logic, or neural nets, among others, all of which would deal with the problems of individual variances at the data input or user levels.

Treatment spectrum: Treatment interventions are perhaps the most familiar. These are the actions taken to identify and correct current problems. Treatment may be seen as a failure of prevention which then requires more expensive interventions in order to return the patient (data) to a more acceptable status. While not all disease is preventable (just as not all data problems are avoidable), many are and the goal should be to anticipate and prevent problems from reaching the treatment stage. Again, the treatment interventions range across a spectrum.

Case identification encompasses those efforts needed to detect problems. Medically this occurs when the patient sees a doctor with a complaint and the doctor responds with a battery of tests, examinations, consultations, and interviews all of which produce not cure but diagnosis. Similarly, organizational data problems must also be identified and diagnosed for cause before they can be addressed with corrective actions. In the data quality management field, case identification can range from proactive (very nearly preventive in nature) to extremely reactive. Proactive case identification includes help desk call analysis, automated edits, automated metrics, ongoing standards based data audits, benchmark analysis, surveys, and use of flexible analysis tools. In the best designed data quality systems proactive case identification plays an key role. Problems discovered as a result of proactive case identification (e.g., analysis of help desk calls) can provide early identification of real problems. case identification process information, routinely compiled for regular analysis can pinpoint incipient or emerging problems and allow managers to implement corrective actions. Reactive case identification occurs at the other extreme of the case identification spectrum. Reactive case identification is stuff of headlines and corporate firings: "Millions of social security checks issued in error (film at eleven)". Reactive case identification occurs when a critical data customer notices something wrong. While corrective actions associated with proactive identification are inexpensive and provide adequate lead time for correction, reactive case identification is, by definition, a crisis situation – it is expensive, time constrained and high pressure. At the extremes it can threaten the survival of the enterprise. Between the extremes of proactive and reactive case identification lies a gamut of situations.

Standard treatments for known disorders is the logical follow on to case identification within the treatment spectrum. In the medical arena this includes the standard array of surgeries, medications, and therapies. In an information organization this may include targeted training efforts, additional automated data edits, reprogramming efforts, etc., ranging from minor corrections to major projects with associated high costs.

Maintenance interventions: The final area of the intervention spectrum is that of maintenance which is further subdivided into two sub-areas: "compliance with long term treatment" and "after care / rehabilitation." The thrust here is to ensure that previously discovered problems and the interventions taken to correct them are continued in force so as to prevent reoccurrence. Medical examples obviously include long term medications (arthritis pills, insulin for diabetics, etc.) as well as regimens needed to regain functionality temporarily lost as a result of disease (physical therapy for stroke, etc.). The analogy is not as strong for information data quality management but should involve continuing refresher training for specific groups (data producers, users, customers, etc.), periodic focused audits, and the like. Technology enhancements may also (arguably) fit into this range. For example, just as newly developed drugs may be substituted for older, less effective drugs in a medical treatment regime, so may software or hardware upgrades be introduced to improve the ongoing maintenance of systems.

PREVENTIVE LEVERAGE IN ACHIEVING DESIRABLE OUTCOMES

Following on to the IOM "Range Of Interventions" model must come a further understanding of the competitive leverage provided by a prevention based approach to managing data quality. Figure 2 is, again, an attempt to illustrate these principles. The adage "an ounce of prevention is worth a pound of cure" is as true in the data quality business as it is in the medical business. Medical and epidemiological research provides sound financial justification for the positive return on investment of investing in prevention. Referring to the IOM model, the data strongly supports investment in universal interventions (immunization, inoculation, sanitation, etc.) as well as most selective interventions (mammography, sickle cell screening, etc.). While it is less clearly established in the still emergent area of individual interventions the data nonetheless supports the thesis that prevention saves money and enhances the quality of life. The state of research in the data quality management field has yet to substantiate or quantify similar benefits in that arena however, it seems entirely reasonable to presume that they exist in some proportion.

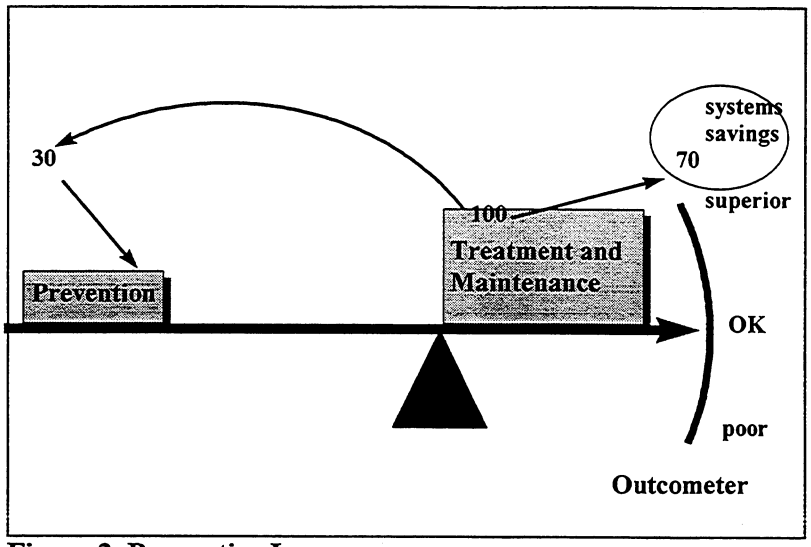

Figure 2 Preventive Leverage

As figure 2 depicts, most systems exist in a state of equilibrium. Generally speaking a successful organization is achieving the ongoing outcomes (profitability) necessary for continuing survival. This organizational stasis is achieved over time by balancing the costs needed to prevent problems with the resources needed to correct problems. From an organizational management perspective a solid argument can be made that, up to a point, resources directed toward risk factor identification and risk mitigation will produce a net savings to the organization by reducing the resources needed to correct product defects. These savings can then be used to either improve outcomes (quality) and gain further competitive advantage, or returned as profits. The downside of course is that in this state of organizational equilibrium all current resources are accounted for in order to sustain the present level of outcome quality. Thus the data quality manager must either restructure his own budget or convince upper management to free up the additional resources needed to implement a program of risk based prevention. Solid evidence for the existence of these benefits exists in the history of the automotive industry in America as they quite literally fought and won the war for their continued existence on the battlefield of product quality. In the case of the auto industry the approach taken to achieve quality gains was based on the "Total Quality" precepts advanced by Edward Deming which also emphasizes upstream defect prevention.

IDENTIFYING ENVIRONMENTAL RISK FACTORS

Having established the case for a risk based, prevention oriented approach to achieving data quality it is now time to develop a practical methodology for implementation. As a minimum, any tool should meet the following criteria:

- Ease of use
- Wide applicability to data systems

- Structured approach
- True "systems orientation" (i.e. encompassing the enterprise wide spectrum)

In addition, a successful tool will achieve the following results:

- Identify significant risk factors present in the organization (risk factors being those things which predispose the system to producing poor quality data)
- prioritize potential interventions in terms of their potential impact on improving data quality

Borrowing from the field of epidemiology, Developmental Research and Programs, Inc. (DRP)[vii] of Seattle Washington has developed a risk focused prevention strategy that has been successfully used to reduce adolescent problem behaviors including substance abuse, delinquency, and violence within communities. Faced with a common problem of overwhelming need and limited resources, DRP has developed a structured approach based on sound science that enables community leaders to analyze their environment, identify the factors that increase the likelihood of adolescents engaging in problem behaviors, and develop programs to concentrate scarce community resources to address these known risk factors. The DRP tool, which we have adapted and extended to address data quality improvement efforts, takes precisely the same approach to data quality improvement.

THE DRP MODEL

Figure 3 displays, in greatly summarized form, the essence of the DRP risk factor model. The model began with the identification of adolescent health and problem behaviors. The five problem behaviors addressed in the DRP model include substance abuse, delinquency, teen pregnancy, school dropout and violence. Reduction of these problem

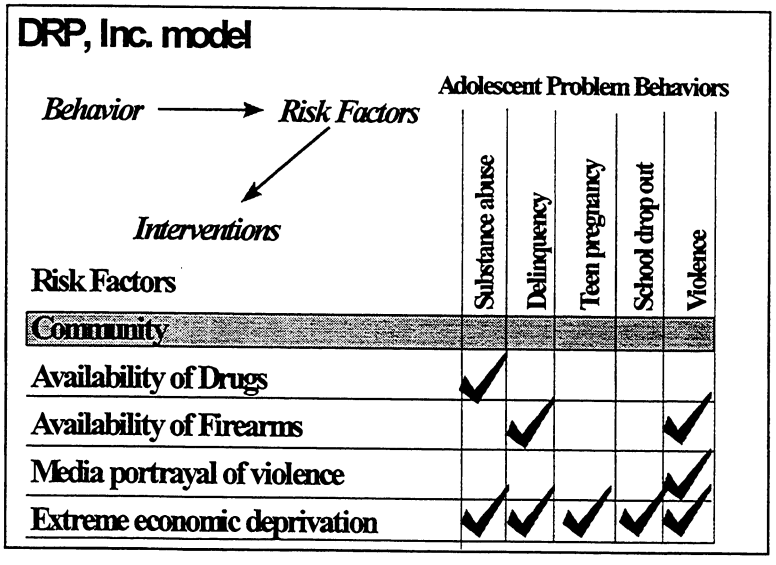

Figure 3 A model for linking problems to environmental risk factors

behaviors is the objective of the program. As a next step, DRPs' analysis of current research identified 19 risk factors, or predictors, for these five problem behaviors. These risk factors are clustered into four domains, or spheres of influence, in a young person's life: community, family, school, and individual/peer. Figure 3 is a truncated matrix for one of these domains (in this case the community). Collecting and analyzing data that are indicative of the levels of prevalence for each of the 19 risk factors within a given community allows a community to select priority risk factors that, if mitigated, will achieve the greatest possible impact on problem behaviors. From the information in these matrices the community is then able to focus on which risk factors must be mitigated in order to achieve the greatest possible impact on reducing problem behaviors. From this example you can see that depending upon available resources and community desires, the city can begin to formulate a coherent strategy to reduce problem behaviors through the mitigation of priority risk factors based on the data. In essence, the tool makes preventive intervention less a "shot in the dark" and more a targeted effort.

EXTENDING THE MODEL: A DATA QUALITY MANAGEMENT CASE STUDY

This case study involved the implementation of a new source data collection system intended to gather information about outpatient visits from over eighty Air Force hospitals and clinics world wide. The implementation took place over a two year period and involved not only the design, planning and implementation of a new data system throughout the enterprise, but also required significant changes to existing business practices, as well as training and education in new skill sets for the entire clinical staff. In the course of implementation it became apparent that the impact of the new system was significantly broader than originally anticipated. Senior leaders, reacting to a ground swell of negative reactions from the field, expressed grave concern. While they strongly supported the system and deemed the information being

Risk Focused Data Quality Analysis Methodology

1. ***Identify critical communities***
2. ***Identify problem behaviors***
3. ***Identify risk factors***
   - ***Conditions that predispose:***
     - ***Universal (Organizational)***
     - ***Selective (Sub populations)***
     - ***Individual (personal)***
4. ***Matrix problem behaviors to risk factors***
   - ***Stratify by community***
   - ***Apply severity scoring scale to problems***
5. ***Rank order problem behaviors by score***
   - ***Select highest priority problems for intervention***
6. ***Design specific interventions***
   - ***Align along intervention spectrum***
   - ***Give priority to actions in prevention range***

Figure 4 Model for risk focused analysis

gathered as critical to future survival, they were also deeply concerned about the negative reaction to its implementation. More specifically, they feared that if user concerns were not resolved the quality of the data being delivered would be inadequate to satisfy the business needs. Our guidance from the senior leaders was to identify the principle problems and recommend solutions.

While using the same basic approach, it was necessary to fine tune the DRP model in order to address our specific need to identify and prioritize risk factors within the organization. The primary adaptation was the addition of a subjective "severity scale" to the problem behaviors matrix. The steps in this approach are listed in figure 4 and described in greater detail below

Step 1. Identify critical communities.

Our definition of critical communities in the DQ process paralleled that suggested by Strong, Lee, and Wang in their paper "Data Quality in Context"[viii]. In the paper they proposed the concept of a data manufacturing system with a data production system that transforms data into useful information. Within this process they identified three roles; that of data producers (those who generate data), data custodians (those who provide and manage computing resources), and data consumers (those who use data). We identified data producers, data customers and data overseers, essentially the same groups.

Step 2. Identify problem behaviors

Departing from the DRP model which defines a common set of problem behaviors for all groups, we then identified generic "problem behaviors" unique to each critical community. Surprisingly, although we had expected to generate a large list of such behaviors, we found that the "problem behaviors" were relatively few. Although many candidates were initially suggested, most of these behaviors were identified as variations on a theme and were eventually reduced to a relative few. These problem behaviors as well as the associated critical communities are displayed in figure 5.

CRITICAL COMMUNITY	PROBLEM BEHAVIORS
Data producers	Inaccurate data entry Delayed (late or no) data entry Incomplete data entry
Data custodians	Failure to recognize problems Failure to measure DQ performance Ineffective/inefficient problem resolution
Data Consumers	Don't use data Apply data incorrectly / inconsistently

Figure 5 Critical data communities and problem behaviors

Step 3. Identify risk factors

In this exercise we attempted to identify those conditions in the business environment that predisposed any of the members of the three critical communities towards the problem behavior. This was accomplished informally by brainstorming and surveys within the critical communities. We were able to improve the thinking process further by structuring the risk factor analysis along the lines of universal (organization wide), selective (peculiar to a specific subpopulation), and individual (personal) factors as suggested by the IOM model. We also found the work of D. Strong, et. al. , helpful . In their paper, "Beyond Accuracy: What Data Quality Means to Data Consumers"[ix] they had previously identified 179 separate "attributes" of data quality. Although the comprehensive list is not provided in this paper these "attributes" encompass items such as: relevance, ease of use, timeliness, level of standardization, clear data responsibility, etc. Upon review we found that most of these "attributes" fit our definition of a risk factor. That is, their absence in the data production system increases the risk that poor quality data will result.

Risk Factor Analysis
Critical Community: Data Producers

Behavior → *Risk Factors* / *Interventions* / Risk Factors	Producer Problem Behaviors: Inaccurate entry	Untimely entry	Incomplete entry	TOTAL	Priority
Data Producer					
No clear standards	2	3	3	8	5
Conflicting incentives	3	5	5	13	2
Low skills in coding	4	4	4	12	3
Low management interest	2	4	4	10	4
Perceived as irrelevant to job	4	5	5	14	1

Figure 6 Data Producer Risk Analysis

Risk Factor Analysis
Critical Community: Data Consumers

Behavior → *Risk Factors* / *Interventions* / Risk Factors	Customer Problem Behaviors: Don't use data	Use incorrectly	TOTAL	Priority
Data Consumers				
Poor access through AIS	5	3	8	1
Don't understand data	1	5	6	2
Data outdated	1	1	2	
Not relevant to job	4	0	4	3
Data incomplete	1	1	2	

Figure 7 Data Consumer Risk Analysis

Step 4. Matrix risk factors to problem behaviors

As in the DRP model, we created a matrix for each of our critical communities. Unlike the DRP model however we had no specific research to support the impact of risk factors on problem behaviors. Lacking this we opted to have critical community members and other key organizational personnel subjectively score each risk factor on a 1-5 scale, rating its correlation with producing the problem behaviors. Figures 6 and 7 provides a summarized version of the matrices for two of the three critical

communities (data producers and data consumers. The matrix for data custodians is not shown) with scores for the top risk factors included.

Step 5. Rank order risk factors by score

Finally the scores were totaled for each risk factor and rank ordered by score with higher scores indicating a greater perceived negative impact on the data production process. The result was a prioritized list of high risk factors currently existing in the organization's data production environment that had a negative impact on quality. Of particular note, these factors reflect the opinions of the primary clients within each critical community (data producers, data custodians, data consumers). Restating this important point, the list tells us what the data producers see as the primary system obstacle to data quality on the input side of the system. Likewise we have prioritized lists from the data custodians and the data consumers. This then is the raw material which is now translated into targeted corrective actions.

Step 6. Design specific interventions

Finally, the process culminated with the design of specific interventions explicitly tailored to mitigate the risk factors identified as a result of the scoring process. Interventions can include such things as training programs, policy changes, software redesign, user interface changes, revised performance incentive programs, or business process changes, among others. Two points need to be made regarding the process of determining what interventions are needed and how they are to be designed.

First is the subjective nature of the determination. To this point the process has provided a structured, step by step approach to aid in determining what deficiencies (environmental risk factors) exist in the data quality process and the relative contribution of each of these deficiencies to the total DQ shortfall. To the extent that the process is followed, it provides a relative index of severity for each of the problems in the data production process as perceived by the producers, users, and custodians. From this point it is a managerial decision as to which risk factors are the most severe and which can be addressed within the political, financial, logistical, and time constraints faced by the enterprise. For example, the model may suggest that a primary causal factor for DQ problems is a lack of specific upper management policy guidance that causes confusion regarding expectations. Management may however make a reasoned judgment to devote resources to immediately mitigate a lower priority risk factor (such as providing additional training) that can be addressed within a short period of time, and defer needed policy changes until the political environment allows. Likewise, a choice could be made to mitigate

multiple small risk factors rather than expend all available resources on a single major risk factor. Assessing the relative impact of the rank ordered risk factors on the total enterprise mission is an inherent responsibility of management. The risk based assessment process outlined in this paper is simply a means to provide a structured, philosophically cohesive construct within and against which management judgment can be applied.

A second point is the form that the interventions should take. Once management has decided which of the rank ordered "risk factors" is to be addressed, it must then decide how this is to be accomplished. The rules of thumb for designing interventions then are to:

- Align the desired interventions along the IOM "range of interventions" spectrum. Risk mitigation programs can take many forms depending upon the problems to be addressed. Urgent extant problems causing significant current damage require "treatment" interventions. Once the current problems are resolved, interventions may then move into the "maintenance" mode. By contrast incipient, potential, or less severe risks may indicate the need for more purely "prevention" oriented options. In reality, every data system will require a mix of risk mitigation strategies in order to achieve acceptable DQ performance. The management task is to determine which strategy is most appropriate to the situation.

- Give priority to programs in the prevention range. Recall the principles established in the IOM "range of interventions" model: Interventions that are most effective, enduring, and low cost are those that are farthest to the left of the range scale, i.e., those in the "prevention" range as opposed to "treatment". Prevention is always cheaper than treatment. Thus, when selecting and designing risk mitigation strategies, priority should always be given to prevention over treatment.

Figure 6 depicts the final results of our analysis of risk factors within the subject program along with the risk mitigation actions proposed. The results are shown as they were presented to executive management and are displayed in the form of a "quadrant analysis" chart (the preferred format for presentations to our executive management). For the purposes of this paper the items of particular importance are contained on the lower half of the chart. The lower left quadrant lists the major deficiencies revealed as a result of the risk analysis exercise while the lower right quadrant shows the strategies proposed to address these deficiencies. It should be noted that these listings are intentionally broad and are intended to present the senior organizational leaders with a "high level view" of the nature of problems causing the undesired effects. Each of the individual items listed can be further decomposed into more specific action items. As of this writing the implementation of these recommendations is actively underway. Because of the incomplete and ongoing nature of the corrective actions there has not yet been sufficient quantitative data generated to confirm or reject the effectiveness of these actions or the validity of the methodology. Despite the lack of quantitative data however, the conclusions and recommendations generated by the model have been subjectively validated by other independent internal program evaluations.

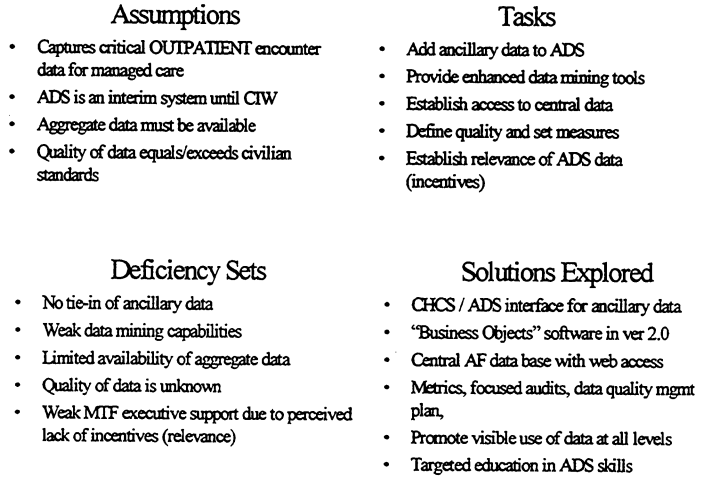

Figure 8 Quadrant Analysis: Recommended High Leverage DQ Actions

CONCLUSION: FINAL THOUGHTS

As with many data managers, our work began with a very practical requirement: the need to understand the cause of a series of problems related to a specific automated information system and then design and implement actions to correct those problems. What we quickly discovered was that while the symptoms of poor quality were readily visible, the causes were most often subtle, elusive, and pervasive. Quick patches often made the immediate problem seem to go away, only to reappear again in a different form at a later date. Ultimately, the quick patches only added an additional layer of complexity to an already convoluted system and buried the real "root causes" even deeper in the organizational structure.

Over the two year period of time during which we dealt with the system that is the subject of this case study, we came to understand that an automated information system is best viewed as an integral component of the larger enterprise which it supports, and not, as is often done, in isolation. It was perhaps a fortuitous side benefit of working in the medical field that we became familiar with the developments, models, and techniques in the field of epidemiology and were immediately struck with their potential application to data systems management. Although viewing data systems as "living organisms" may appear unusual, it is my firm belief that the constructs, models, and tools which we have described in this paper are almost universally applicable across the data systems management field.

Although our application was limited to addressing data problems within an existing (although newly deployed) system, I believe the model is equally useful in many stages of the system life cycle.

- In new systems the "range of interventions" and "critical communities" models provide a comprehensive "check list" of critical design characteristics needed in a successful system. By ensuring that these considerations are a part of the initial design process a system can be constructed with quality that is "built in" and doesn't need to be "added on" at a later date (and at added expense). Just as the formal system life cycle development process model adds discipline to the new system design process, so do these models add discipline to thought processes about what we expect from our systems and how we intend to ensure that those expectations are met. By making the prevention, treatment, and maintenance spectrums a part of the conscious thought process, designers can ensure that the infrastructures needed to support these considerations are in place before the system is deployed.

- In existing systems the risk focused prevention approach can provide a structured approach to the diagnosis, definition, prioritization, and treatment of system problems that result in poor data quality. Especially valuable is the ability of these tools to address not just the symptoms of poor quality, but more importantly the root causes that are often so deeply masked. By taking a total systems approach to the problems, the approach can also place data quality problems in perspective for key decision makers, provide a relative priority for problem issue resolution, and give both operational data managers and organizational leaders a quantitative baseline for cost-benefit decisions.

- Finally, I believe that this model has the potential to form the philosophical foundation for a comprehensive formal data quality plan. As a next step in our data quality journey, I see the need to

create a formal organizational data quality program to guide and integrate our efforts across the enterprise. My review of data quality plans and guidance from a variety of organizations, while interesting and educational, has not been totally satisfying. Most of these data quality plans are written from the data processing departments point of view -- not totally surprising given that this has traditionally been the locus of such responsibility in many organizations. However, with recent movement towards organizational "data stewards" and the recognition that information is an enterprise asset with the potential to exert tremendous leverage in the competitive arena a more robust approach is needed. A truly effective data quality program must recognize the pervasive nature of information systems in our organizations and address the diffusion of responsibilities for the quality of the information produced. By placing information systems into context as an integral part of the enterprise the risk focused, epidemiological approach has exceptional promise to do just that.

---

[i] Zwerling, Craig, The Epidemiology of Firearm Deaths in Iowa, 1980, *American Journal of Preventive Medicine*

[ii] Novotny, Thomas E., Smoking Among Black and White Youth: Differences, Nov 96, *Annals of Epidemiology*

[iii] Lenaway, Dennis D., The Epidemiology of School-Related Injuries, May 92, *American Journal of Preventive Medicine*

[iv] Wintemute, Garen J., The Epidemiology of Drowning in Adulthood, Nov 88, *American Journal of Preventive Medicine*

[v] Collins, James J., Epidemiology of Alcohol-Related Violence, 1993, *Alcohol Health and Research World*

[vi] Institute of Medicine (IOM), 1994, Reducing Risks for Mental Disorders: Frontiers for Preventive Intervention Research. Matricia J Merzed & Robert J. Haggerty, Eds. Washington, DC: National Academy Press

[vii] Developmental Research and Programs, Inc., 130 Nickerson Suite 107, Seattle, WA, 98109. Communities that Care®, a community training system for risk-focused prevention.

[viii] Strong, D. M., Lee, Y. W., and Wang, R. Y., Data Quality in Context, *Communications of the ACM,* Forthcoming

[ix] Strong, D. M., and Wang, R. Y., Spring 1996, Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems,* Vol. 12, No. 4, pp 5-34