

Reusing Clinical Protocol Content to Improve R&D Productivity

Protocols are instructional manuals for clinical research, describing study procedures and their rationale and serving as a guide for how study results will be interpreted and used. Every clinical study begins with a protocol, and all start-up, conduct, and reporting activities refer back to it. Therefore, the clinical protocol may be viewed as centrally located in a clinical research knowledge network, whose nodes represent study personnel and their means of conveying knowledge, and whose linkages represent knowledge transfer among personnel. Document-centric practices have limited the efficiency of clinical

knowledge transfer by discouraging computer-assisted content reuse. In order for computers to facilitate content reuse, some content must be structured and semantically modeled. Because of its dominant network centrality, the structured protocol environment is ideally situated to become the primary conveyance for new knowledge about planned and in-progress studies (ie, the study metadata). Enterprise-scale computer-facilitated reuse of protocol-sourced content is predicted to have near-immediate measurable benefits on study conduct quality and operational efficiency.

Fredric J. Cohen, MD

Medidata Solutions
Worldwide Inc,
Conshohocken, Pennsylvania

Key Words

Content reuse; Extensible protocol; Knowledge transfer; Productivity

Correspondence Address

Fredric J. Cohen, Pharma
Growth Strategies, LLC, 13
Summit Center Square, Suite
132, Langhorne, PA 19047
(email:
fred@pharmagrowth.com).

Presented at the 21st Annual
DIA Conference for
Electronic Data
Management, February
2008, Philadelphia,
Pennsylvania.

THE CLINICAL RESEARCH KNOWLEDGE NETWORK

Clinical research is essentially the creation and dissemination of knowledge. Clinical protocols are instructional manuals for clinical research, describing study procedures and their rationale and serving as a guide for how study results will be interpreted and used (1). Every clinical study begins with a protocol, and all start-up, conduct, and reporting activities refer back to it. Therefore, contributors to the clinical protocol may be viewed as centrally located in a multiorganizational knowledge network (the clinical research knowledge network) with nodes representing study personnel and their means of conveying knowledge and linkages representing knowledge transfer (2) (Figure 1).

When the research project's work outputs are considered in light of the prerequisite knowledge permitting those activities, some basic principles underlying this knowledge network become apparent.

First, knowledge transfer must be thoughtfully time-coordinated among diverse functional experts. Delays in creating, transferring, or using knowledge that is needed to begin or to com-

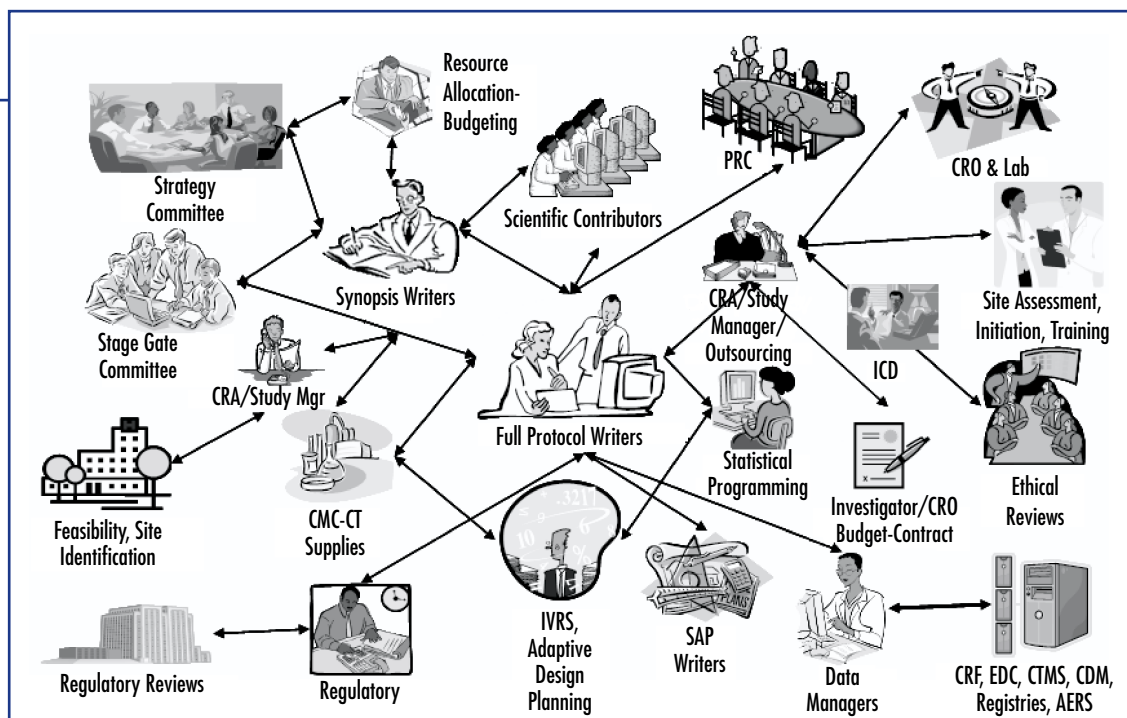
plete an activity can have profound effects, not only on the activity itself but also on other activities that are dependent upon it. Viewed this way, knowledge management is not simply an effort to promote organizational flexibility and innovation but is also an efficiency enhancement effort, akin to other enterprise-scale endeavors that seek to measure and improve organizational business processes, such as enterprise resource planning (3,4).

Second, if the centrality of protocol-sourced information is accepted, it follows that knowledge must be transferred between protocol authors and other internal (to the research organization) functional experts and external contributors and end users efficiently and effectively for the study to start and to run efficiently and effectively.

Third, although the mechanisms of knowledge transfer and content reuse differ according to functional area and task, standard practice today is for people to formally (via long-form documents) and informally (via email threads, meetings, etc) transfer knowledge about the study and its resulting data (ie, its design, rationale, methods, etc) without intellectual support from computers. Rather, it is typical today for computers to provide the equivalent of ad-

FIGURE 1

High-level view of the clinical research knowledge network during study start-up. Shown are a fraction of the people and information conveyances needed to start and operate a typical clinical study. PRC, protocol review committee; CRO, contract research organization; CRA, clinical research administrator; ICD, informed consent document; CMC, chemistry, manufacturing, and controls; CT, clinical trial; IVRS, interactive voice response system; SAP, statistical analysis plan; CRF, case report form; EDC, electronic data capture; CTMS, clinical trial management system; CDM, clinical data management; AERS, adverse event reporting system.



ministrative assistance for knowledge workers, such as the following:

- Word processors and related applications (eg, spreadsheets) shape the form of conveyed information.
- Email speeds dissemination of information.
- Collaboration portals encourage explicit communication.
- Document management systems provide efficient storage and security of written information.

Computers have not typically helped researchers communicate information unambiguously; information must be reinterpreted at every information use and reuse juncture. They also have not been particularly useful for interrelating information that spans functional areas (knowledge domains), necessitating constant supervision of newly created documents by knowledge domain experts. Because of these two limitations, computers have also played only a supporting role in propagating knowledge over time—transferring knowledge from domain experts to novices—the institutionalization of which is the basis of organizational memory (5,6).

It is hypothesized that for the clinical research knowledge network to operate most efficiently,

the power of computers must be harnessed to operate in the areas of knowledge transfer where it does not today.

KNOWLEDGE TRANSFER INEFFICIENCIES IN THE DOCUMENT-CENTRIC CLINICAL RESEARCH ENVIRONMENT

The contemporary clinical research environment is strongly document-centric. Preformatted documents, such as text reports, slide presentations, spreadsheets, and so on, are used as the primary conveyances of research knowledge. Substantial effort is spent specifying document formats, creating documents, and storing and reusing documents to retrieve, use, and reuse the information contained within them.

Information technologies as used today facilitate more geographically dispersed and more complex studies than were feasible previously but, as alluded to above, they fall short in terms of using and reusing information efficiently and effectively. No matter how quickly documents can be distributed among research personnel, when they reach their intended targets, they must be read and interpreted in order to be used. Therefore, unstructured documents allow

the original intent and meaning of their content to differ from one person to the next. As a result, efforts must be expended to discover the intentions of authors and to correct mistakes made when storing, retrieving, interpreting, reusing, and implementing information found in documents. These information discovery and error correction activities are not value adding; that is, they do not contribute positively to work output or its value. It follows that reducing or eliminating them will result in improved return on investment and productivity gains by freeing resources.

Opportunities for improvement in this area are readily apparent when the repetitive activities necessary for formal information transfer and reuse are represented as a continuous cycle—the protocol information reuse cycle. At its most basic, the cycle of protocol information use and reuse can be described by five steps (Figure 2):

1. Information creation
2. Information storage
3. Information retrieval
4. Information interpretation
5. Information reuse

At each step in the information reuse cycle lies an opportunity for information to leak from the cycle, that is, for the transfer of knowledge to be interrupted or corrupted, such that work is required to resume the cycle. Note that the concept of information leak as used here is different from more common usages referring to intentional or unintentional disclosure of proprietary data or to the loss of key information-bearing employees. Definitions and examples of how information leaks from each step in the protocol information reuse cycle follow.

STEP 1: INFORMATION CREATION

The de novo creation of concepts that will be used to convey knowledge is information creation. In the document-centric environment, this information is created and conveyed in the context of documents, such as the research protocol. The cycle occurs when reused information (from step 5) is merged with de novo infor-

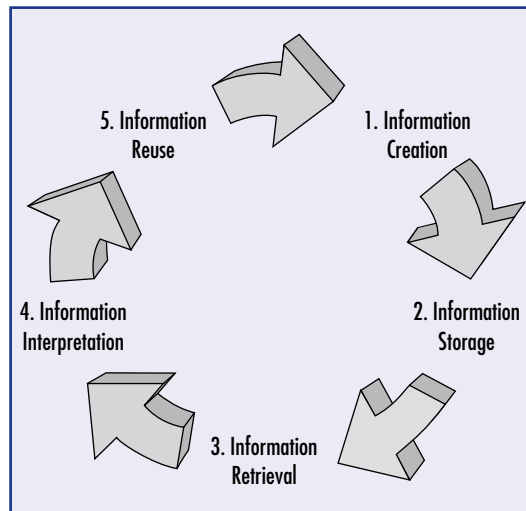


FIGURE 2

The information reuse cycle in clinical research. See text for explanation.

mation (step 1), for example, to create a new protocol or a protocol-dependent document.

An example of information leak at this step is the operational ambiguity created when one part of a document says one thing and another says the same thing in a different way or contradicts the first instance (7). Ambiguous language is a leak of information in the sense that readers of documents containing ambiguities require additional information to act on the documents as intended or else risk acting as not intended. Therefore, information leaks but is recoverable with additional work.

Another common example of leakage at step 1 is the loss of rationale for certain choices made by authors that are difficult to ascertain from reading documents. Rationale is an easy target for noninclusion by document editors. For example, in the United States, inclusion of rationales for study design or procedure choices is not codified by statute (see 21 CFR 312.23), leaving subjective judgment and sparse therapeutic area-specific regulatory guidance as the only means of determining whether inclusion of a study design rationale is necessary.

STEP 2: INFORMATION STORAGE

Step 2 is the storage of the information created in step 1. Storage requires identifying the definitive instance of information (ie, the latest document version) to be stored. This is not trivial when multiple authors are involved, particularly

when a document management system is not used to manage the authoring workflow, as is common practice for protocols.

STEP 3: INFORMATION RETRIEVAL

Information stored in step 2 is physically retrieved in step 3. Such retrieval is not a straightforward task when the repository of information is large, as it is for large pharmaceutical or medical device companies that conduct many trials simultaneously. Information retrieval often requires a text search of keywords to find relevant material. Without semantic (ie, meaning-based) search capabilities, such searches are often inefficient, with a high return of irrelevant information. For the same reason, information conveyed in a document cannot easily be combined conceptually with information stored in related documents.

STEP 4: INFORMATION INTERPRETATION

Following retrieval, document-conveyed information must be interpreted in order to extract meaning from it. Ideally, the intention of the original authors will be transmitted to readers (users).

Leaks at this step are most readily appreciated by example. Consider a clinical protocol whose schedule of activities includes the task “routine exam.” Is this an initial exam or a follow-up? Is it brief or comprehensive? Does it include a detailed history, a brief history, or no history? The reader cannot make these determinations from the term “routine exam” alone, necessitating additional work to determine the original authors’ intentions or else risk acting as the authors did not intend.

Consider also the need to transmit document-conveyed information to dependent information systems, for example, the transfer of protocol information to the electronic data capture (EDC) system. It is not unusual for nonauthors of protocols to program the EDC system, adding a layer of communication to the interpretation of protocol information prior to its reuse.

STEP 5: INFORMATION REUSE

Information is reused when original concepts, though not necessarily original text, created at

step 1 are employed for either the same or a new purpose. In the case of the protocol, information is frequently reused when creating new protocols and protocol-dependent documents, such as the case report form, investigator brochure, study report, and so on.

Information reuse is typically manually controlled via document-to-document transmission, using desktop office tools like copy/paste and “file save as.” These processes allow human errors in process execution and judgment to cause some information loss and corruption.

This type of leak also occurs when information systems reuse information created upstream. Using the example of the transfer of protocol information to an EDC system again, the protocol information must not only be interpreted, it must also be recoded from English or some other written natural language into a structured language, which can be understood by the EDC system to allow its transfer.

IMPROVED KNOWLEDGE TRANSFER EFFICIENCY WITH SEMANTICALLY MODELED STRUCTURED CONTENT

Reducing information leaks during information reuse will necessarily reduce the activities needed to correct or compensate for them, potentially leading to productivity improvements. Relatively recent advances in computer science have created an opportunity for computers to play a more direct role in reducing such leaks.

The information tools described earlier, such as word processors, have always been capable of transmitting unstructured content in its native form, what is usually referred to as natural language, with reasonably high fidelity. What these commonly used tools have not been capable of doing until more recently is transmitting the structure of information along with its form. Here, the term “structure” refers to the components of information and their interrelationships that, together with form and grammatical conventions, give language meaning (8).

Natural language information conveyed in documents has an implicit structure, by virtue of its language rules and form, that is readily interpretable and analyzable by humans. Comput-

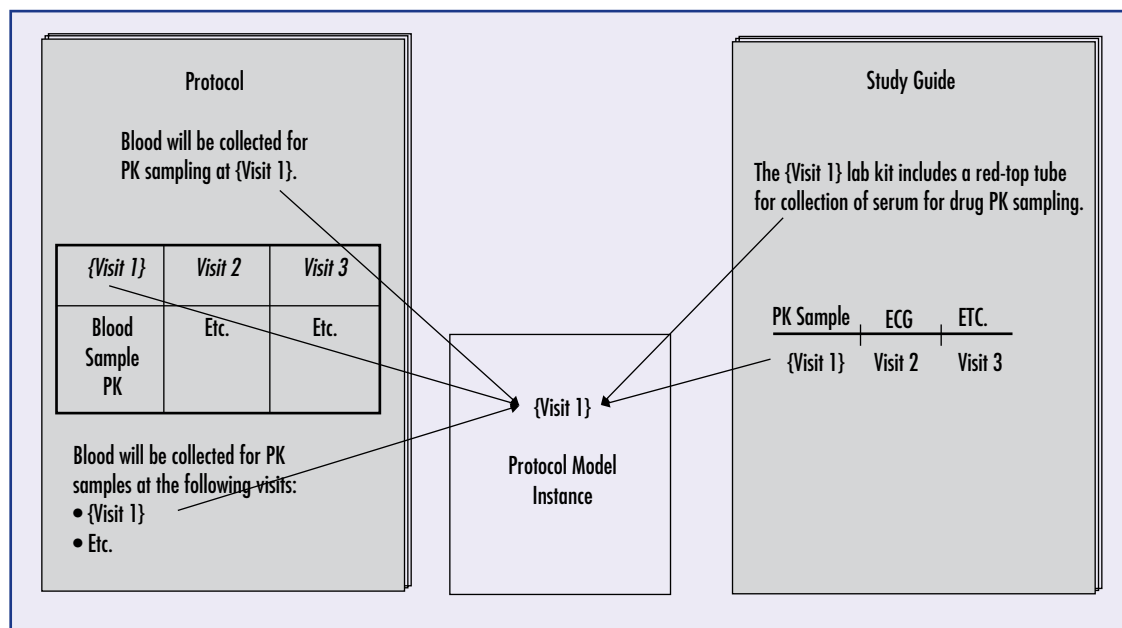


FIGURE 3

Reuse of extensible protocol content. Idealized view of document content reuse in two related documents, the research protocol and study guide. The data element named Visit 1 is given meaning within the context of a protocol knowledge model. The Visit 1 concept may be accessed from a single protocol model instance and reused in multiple documents and information systems multiple times, through human-directed or automated data transfer, while preserving its authors' original meaning and usage intent. PK, pharmacokinetics.

ers, conversely, have historically not been able to infer the structure of information from its form or content alone. Therefore, in order to improve computer facilitation of knowledge transfer, it was necessary for information scientists to create so-called markup language conventions that confer explicit structures to information components. Once an explicit structure was present, computers could assist in the interpretation and analysis of document content, much as they had historically assisted with analysis of numerical and character data collected during research studies, which were stored in and accessed from relational databases. Documents whose key concepts are represented within a formal hierarchical structure can be considered to be semantically modeled, in that the concepts have been abstracted in a manner that confers meaning on them (9). If implemented ideally, such meaning will be highly domain relevant and unambiguous to domain specialists.

Structured information does not usually contain the nuances and idiosyncrasies that make natural language a delight to explore creatively, but being delightful is not a required property of research documents. And there are no restrictions against using natural language

around structured information in a document to provide readers with enhanced context and readability. As a trade-off for time spent creating mixed structured and unstructured documents, maintaining and transmitting structured information with computer assistance reduces risks of information degradation and loss over time, space, and knowledge domain.

THE EXTENSIBLE PROTOCOL

The information interpretation step (step 4) is particularly amenable to computer facilitation. Indeed, if a research protocol is sufficiently information dense, and all relevant information is structured, use of appropriate tools could theoretically eliminate the need for human interpretation of key concepts among authors and users, reducing the five-step reuse cycle to a four-step cycle (Figure 3).

The benefits of reducing or eliminating information interpretation become self-evident when considering the transfer of information from the protocol (via protocol authors) to downstream systems such as EDC. In the case of EDC setup specifically, a clinical data domain expert is usually asked to interpret the protocol before coding the EDC system to create the nec-

essary data capture modules. Instead, if a tool were used to create a semantically modeled protocol and the protocol tool and EDC tool could communicate with each other, no human interpretation by the data domain expert or anyone else would be needed to set up the EDC system. It would be necessary only to transfer (or clone) the structured protocol data into the EDC system. We use the term “extensible protocol” to describe such semantically modeled protocols, owing to their intrinsic ability to extend the utility of information created for the purpose of preparing a protocol document beyond the protocol document itself.

Practically, the automated transfer of protocol information downstream to other systems and documents is made possible when both an extensible-protocol authoring tool and a downstream information system or document-authoring tool speak the same structured language. The most likely candidate for a universally spoken structured language is XML (extensible markup language—a general purpose specification for creating structured language) operating within rules imposed by a universally accepted data standard. An open-source data standard for clinical research has been promulgated and supported by the Clinical Data Interchange Standards Consortium (CDISC) organization. The specific CDISC standard governing the interchange of protocol data among information systems is known as ODM (operational data model) (10). Conceptually, ODM can be considered an XML representation of an annotated case report form (CRF), which itself is simply a blank CRF that documents the location of the data with the corresponding names of the data sets and the names of variables included in them (11,12).

Note that the above-described benefit of extensible protocols to interchange data with downstream systems is theoretical only in the sense that the magnitude of net benefit that can be achieved is unknown when such an interchange is performed routinely on the enterprise scale. Proof-of-concept demonstrations of this interchange have been performed publicly at several CDISC-organized events, whereby a sin-

gle extensible protocol authoring tool transferred data via an ODM-formatted XML file to several EDC system vendors simultaneously, and these data populated relevant data capture screens.

In addition to greatly reducing the impact of or perhaps eliminating the information interpretation step, routine use of extensible protocol tools could also improve knowledge transfer at all other steps of the information reuse cycle, as described below.

At information creation (step 1), structured information, characterized by its explicitly defined conceptual relationships, can be subjected to logical data checks that can greatly reduce operational ambiguities (7). For instance, a computer could check whether each study objective has a corresponding outcome measure and associated activities in the schedule of events. Such checks could, in large part, supplant human quality assurance of protocol content.

At the storage step (step 2), the definitive instance of a concept can be stored within the working version of a protocol model at all times, ensuring that authors are always working with the definitive version. Furthermore, variably restrictive controls on content can be instituted (eg, locking content by concept area), further reducing leaks at this step. For the first time, it also becomes practical for large organizations to determine the state of protocols-in-process directly (concomitant with the storage of active versions), enabling portfolio tracking at the study planning stage.

At the retrieval step (step 3), use of semantically modeled content provides an opportunity to create semantic knowledge repositories, distinguished from typical document or content repositories by enhanced search and retrieval and analysis functionality. With enhancements made possible by the use of extensible protocols and related content, information separated by time, space, and knowledge domain can be logically associated, searched, and retrieved by users in ways that are intuitive. In the past, such functionality has required the manual population of document-metadata fields to implement.

Document metadata are simply bits of information describing the document, typically accessed in a properties dialog; such metadata include key words, author names, title, and so on. Working within a document-metadata environment, information users find related concepts contained in separate documents by searching on the combination of a concept and a text string.

The work flow is much simpler in a semantic knowledge repository environment, such as that enabled by use of extensible protocols, because metadata richly populate every document automatically as it is created. That is, the study metadata—the numerical and textual data describing the collected study data—themselves become the original source material. Research documents are simply manifestations (archived instances) of study metadata. Study metadata collected from extensible documents may be archived, analyzed, and combined conceptually, independent from—although still referencing—the documents from which they were originally generated, offering novel knowledge propagation and project metrics capabilities.

Finally, structured information reuse is facilitated by automated population of dependent documents and information systems with authoritative source content. Here dependence may be viewed as a time sequence of document or system finalization. Thus, document Y is said to be dependent on document X because document Y cannot be completed before document X is completed. Although some human intervention is needed to perform these tasks, they should be far less susceptible to data corruption or loss than conventional copy/paste and file transfer methods used routinely today, with the use of appropriate controls (ie, automated checks and warnings) and operating procedures.

CONCLUSION

The protocol contains most of the information needed to describe and conduct a clinical research study. When a protocol is sufficiently represented as computer-analyzable—sometimes termed machine-readable—data, its con-

tent may be construed as the definitive source of study metadata. As such metadata are used and referenced continually throughout a trial and are necessary to identify and interpret outcomes from a trial, the protocol and its contributors may be viewed as centrally located in a clinical research knowledge network that influences both the scientific output of a study and its operational efficiency.

The influence of the protocol on study operational efficiency has been recognized for some time (13,14). But it has not been as widely appreciated that the effect of protocols on study operational efficiency depends in part (perhaps in large part) on knowledge transfer efficiencies among the clinical research knowledge network. Specifically, knowledge must be transferred between protocol contributors and authors of protocol-dependent documents and to the data managers and end users of protocol-dependent information systems efficiently in order for a study to operate efficiently. Such knowledge transfer, in turn, depends upon the fidelity of formal and informal modes of information transmission and reuse, because error-prone communications necessarily lead to the non-value-adding activities of locating definitive source information and correcting information errors after they have been incorporated or implemented (eg, protocol amendments, EDC rebuilds).

Representing protocol information as structured, computer-analyzable data—study metadata—that can be exported and shared directly with dependent information systems creates opportunities for markedly improving the fidelity of information transmission and reuse and thus for increasing the efficiency of knowledge transfer throughout the knowledge network. Specifically, reuse efficiency gains are possible at each of the five steps of a typical information reuse cycle. Efficiency gains will be most easily recognizable and measurable at the information interpretation step, which can be disruptively impacted by streaming content from system to system with minimal human intervention, thus avoiding the opportunity for error introductions and rework delays caused by ambiguous

content and normal variability in the interpretation of information among people. Beyond this efficiency gain, it is predicted that all other steps in the information reuse cycle can be impacted by enterprise-scale implementation of this technology. Predicted efficiency gains resulting from its use will be measurable by cycle time reductions and reduced rework activities (eg, protocol amendments). Other predicted benefits, such as improved knowledge propagation and organizational memory, will not be as amenable to short-term measurements but should manifest themselves over time by improved productivity and perceptions of improved organizational effectiveness (15).

Acknowledgment—Medidata Solutions Worldwide provided financial support for this manuscript.

REFERENCES

- Bell HD, Walch KA, Katz SB. "Aristotle's pharmacy": the medical rhetoric of a clinical protocol in the drug development process. *Tech Commun Q*. 2000;9(3):249–269.
- Katz N, Lazer D. Building effective intra-organizational networks: the role of teams. *John F. Kennedy School of Government Center for Public Leadership Working Paper Series* 2003;1:83–107.
- Grant RM. Toward a knowledge-based theory of the firm. *Strategic Manage J*. 1996;17:109–122.
- Newell S, Huang JC, Galliers RD, Pan SL. Implementing enterprise resource planning and knowledge management systems in tandem: fostering efficiency and innovation complementarity. *Inf Organ*. 2003;13(1):25–52.
- Bierly PE, Kessler EH, Christensen EW. Organizational learning, knowledge and wisdom. *J Organ Change Manage*. 2000;13(6):595–618.
- Alavi M, Leidner DE. Knowledge management and knowledge management systems: conceptual foundations and research issues. *MIS Q*. 2001;25(1):107–136.
- Kahn MG. Computable protocol models for interactive trial design. *Drug Inf J*. 2002;36(3):487–497.
- DeRose SJ. Structured information: navigation, access, and control. 1995. Available at <http://xml.coverpages.org/deroseStructure.html>. Accessed March 6, 2008.
- Ontology (information science). Wikipedia. Available at: http://en.wikipedia.org/w/index.php?title=Ontology_%28information_science%29&oldid=195699845. Accessed March 6, 2008.
- Clinical Data Interchange Standards Consortium. Specification for the operational data model (ODM). V.1.3 December 19, 2006. Available at: <http://www.cdisc.org/models/odm/v1.3/final/ODM1-3-0-Final.htm>. Accessed March 6, 2008.
- Souza T, Kush R, Evans JB. Global clinical data interchange standards are here! *Drug Discov Today*. 2007;12(3/4):174–181.
- US Food and Drug Administration. Study data specifications V.1.4 August 1, 2007. Available at: <http://www.fda.gov/cder/regulatory/ersr/StudyData.pdf>. Accessed June 5, 2008.
- Kahn MG, Broverman CA, Wu N, Farnsworth WJ, Manlapaz-Espiritu L. A model-based method for improving protocol quality. *Appl Clin Trials*. 2002;11(5):40–46.
- Zuckerman DS. Trial by design. PharmaExec.com. June 1, 2006. Available at: <http://pharmexec.findpharma.com/pharmexec/article/articleDetail.jsp?id=333303>. Accessed March 7, 2008.
- Jennex ME, Olfman L. Organizational memory/knowledge effects on productivity, a longitudinal study. *Proceedings of the 35th Hawaii International Conference on System Sciences (HICCS '02)*. 2002; (4):1029–1038.

The author has disclosed that he is a consultant to and a stock shareholder in Medidata Solutions Worldwide.