Data Capture Technology:

From Remote Data Entry to Direct Data Capture.

Andy Hyde

Remote data entry (RDE), in the conventional sense has failed. At best, RDE - using computers with electronic forms installed to collect clinical trials data - functions acceptably in a limited number of trials for a limited number of data types. Current estimates predict that no more than 5% of all trials will ever use RDE as the primary method of data collection [1]. But technology cannot be ignored.

Considering the possibilities, we could use our increasing knowledge of computing's strengths and seek a way to capitalize on technology's potential. More and more data-providing systems are based on computers or technology, making it possible to connect a data collection computer to a data provider together and thus to use direct data capture (DDC) to collect data.

A parallel development may have an even greater effect on data collection: the development of hospital systems for collecting patient data from its current equipment and its legacy systems, the electronic patient record (EPR) or electronic health record (EHR). Once a hospital or other health care provider has an EHR or EPR in a database, it becomes possible for clinical trials personnel to extract from it the information they require. This can minimize duplication of data entry - a growing problem at hospitals and investigator sites - and prevent the data quality problems caused by repeated entry and transcription.

Remote Data Entry

Electronic Data Capture (EDC) refers to any method of electronically capturing data for a clinical trial. In practice, RDE - using basic personal computer (PC) technologies - has been used almost exclusively for electronic data capture at clinical trial sites. A typical RDE implementation includes a desktop or laptop PC with an electronic form application installed. An electronic case report form (eCRF) replaces the paper CRF.

The eCRF has several advantages over the traditional paper CRF. The greatest of these is its ability to check data as it is being entered. If an error or omission is detected, the person doing the data entry can be warned so that corrective action can be taken immediately. This obviously has a positive effect on the quality of the data returned to the sponsor. The second greatest benefit is the time saved in transferring the data into the sponsor's central clinical trials database. When this is planned and tested early in the process, data can flow smoothly from one system to the other, reducing the chance of introducing errors into the data. Other benefits include control over additional comments and improved legibility, which can be problems in paper CRFs; and reduce paper usage, a valuable contribution in these days of environmental consciousness.

Yet, after more than 10 years in development, current EDC methods, primarily remote data entry, have been implemented into only an estimated 5% of clinical studies - a percent that is not increasing.

If there is to be a future for RDE, it will be found in an Internet based solution. Several pilot trials are being conducted to examine the feasibility of using Internet technology. Early results show that using ActiveX or Java Applets is currently too slow. Using some kind of server-based data checking with an extremely thin client - it has been called an "anorexic" client - works better. These solutions will, however, still suffer from the same inadequacies of their forebears. All data must be "pre-prepared" and manually entered into the data entry system.

The expectations placed on RDE systems were shortened study duration, reduced errors and resource savings. Only one expectation has been consistently demonstrated; a reduction in error rates in the data returned from the Investigator. Many parts of the process other than RDE affect trial duration. And, so far, little evidence has emerged to demonstrate reduced use of resources - either human or financial.

Direct Data Capture

RDE is just one way of collecting data into a computer. As implemented, RDE requires several steps between source and data entry. But advances in technology are rapidly providing a realistic alternative, direct data capture (DDC). Direct data capture is base on the concept of putting source data directly data collecting computer system.

With RDE all source data is put into one standard format before it is entered into the computer. With DDC, data can arrive in the computer in many different formats; then the computer can manipulate the data to meet one standard for transfer into the central repository.

DDC is not a new concept. Many laboratories already use the method, and many sponsors receive electronic laboratory information that can be imported directly into their clinical databases. The samples are analyzed by a machine which outputs its results in computer form, ready to be re-formatted into a standard layout and transmitted electronically to the sponsor. Once at the sponsor site, the data can be up-loaded into the central repository.

A rapidly growing number of clinical trials data sources are becoming available electronically, for example. Equipment for various aspects of physical examinations - ECG, EEG, magnetic resonance imaging (MRI) and nuclear imaging machines for example - can all create digital output. Image data can be stored in a standard format called DICOM (Digital Imaging Communications in Medicine) - a standard protocol for transmitting medical images, waveforms, and ancillary information. Information about the technical parameters of those machine is stored in the image file as part of this standard and can readily be extracted and transferred into the clinical data store. With tradition paper based CRFs, these kind of technical data have been associated with a high error rate and entering them into an eCRF is time consuming.

DDC is also applicable to non-electronic sources. A large amount of data provided by the research subjects themselves. In RDE-based studies these data are elicited through verbal questioning, written into the CRF and finally entered into the eCRF. With increased computer literacy and ownership, however, many subjects can fill in information at a terminal whilst they wait to see a doctor - or even while with the doctor. In the future, providing information via the Internet before a visit will become an acceptable data collection technique.

"Smart Cards" are another future source of subject information. These optical storage devices, the size of a credit card, can store all of a person's health-related information along with security measures, such as photo identification. The notes taken at each visit can be stored, as can information about drugs prescribed. These cards can hold up to 4 megabytes of compressed information - the equivalent of 2000 A4 or letter-sized pages - enough for a lifetime for most people.

DDC can reduce the number of errors even further than RDE. It can save the Investigator time and effort. It does however require the development of a multitude of different interfaces to the different data providing sources.

A future source of data for DDC is becoming available through technological developments in the health care industry. Health care institutions and the pharmaceutical industry have been facing the same EDC problems. At many health care sites, the solution is the electronic patient record (EPR) and the electronic health record (EHR).

EPRs and EHRs

EPRs and EHRs are an exciting development for information management - a development driven mainly by managed heath care organizations (MCOs) in the United States and by the movement to reduce costs for U.S. insurance companies and government-funded health care in Europe. By combining all the information on patient care for hundreds or thousands of patients receiving a given drug or treatment with treatment outcomes, third-party payers can identify expensive drugs or treatments with poor outcomes and use that cost/benefit information to reduce costs. To combine information on so many patients requires a computerized system, hence the need for EPRs.

There is a slight difference between the EPR and the EHR. The EPR assumes that the person is a patient and is therefore in need of some examination or treatment. The EHR has broader aims: to register the heath information of individuals whether or not they are current patients in need of care. This then extends the information to smoking habits, dietary information, exercise routines and more, including data collected outside the health care institution.

For healthcare institution (hospitals for example) - where implementation of Information Technology has often grown in uncontrolled ways - developing an EPR is a major undertaking. Each department and function autonomously selected a computer

system suited to its own needs without forseeing thatit would one day need to communicate with other systems.

Creating an EPR can be done in two ways. The first is to replace all of a health care provider's systems with one new distributed system based on one platform and one application. But that is a major undertaking, one for which few health care institutions are setup physically, organizationally and politically. This approach may be impossible.

An alternative is to retain all the systems currently in place and get them to "talk" to each other. This is the solution that many vendors choose. A standard was developed for communicating between the systems called HL7 (Health Level 7, in which the 7 indicates that it operates at level 7 of the Open Systems Interface (OSI) 7 layer model for networking). The HL7 standard is approved by ANSI (American National Standards Institute) and is based on consensus of its members. It is possible to add any new data collection system into the EPR as long as it can communicate in HL7.

Technology choices for EHRs

The technology choices made by the healthcare industry should indicate what is acceptable for an investigator's routines and a hospital's functions. Choices thus far have been pragmatic - to avoid changing the daily practice of the caregiver. Data must be collected at the point of care and require no subsequent transcription. In this way, the information is immediately available to other caregivers who may require it. Some of the technologies available are voice recognition, laser disk storage and structured remote data entry.

Voice recognition is currently high on the list. Much of the information collected on a patient visit is recorded to tape, then sent to the medical records department for transcription. "Off-line" dictation of this kind can still be done with digital tape recorders and voice recognition software, or a doctor can dictate directly into a desktop computer. With training (and an American accent in a quiet environment) voice recognition works well. A host of medical dictionary add-ons make it both possible and highly effective.

Laser disk storage. Another technology is computer output to laser disk (COLD) whereby free text is typed into a predefined word processor template document, then "printed" via an interface that can interpret the structure of the form and the text in it. The text and the template are split and the data are written to a mass storage device - in this case, a laser disk. The record can be regenerated by reading the data back into the template.

Structured data entry, such as that for RDE, is far down the list of choices. It requires specialized software and it takes longer to collect the same data than voice recognition, laser disk storage or paper case record forms (CRFs). The greatest chance for the adoption of structured data entry is increases in access to Internet technology, which is becoming a

Translation & Transfer via HL7

HL7 is a messaging system in which messages are structured according to a pre-defined format and sent from one system to another. The sending system needs to know only how to convert its data into an HL7 message and the receiving system needs to know how to extract those data. Message are sent when triggered. The Admissions. Discharges and Transfers (ADT) system for example has a trigger that activates whenever a new patient is admitted to a hospital. The trigger says electronically, "send the name, insurance company, address and other particulars to all other systems." An example of an HL7 message is:

```
MSH|^~\&|ADT1|MCM|LABADT|MCM|1
98808181126|SECURITY|ADT^A01|MS
G00001|P|2.3|<cr>
EVN|A01|198808181123||<cr>
PID|||PATID1234^5^M11||HYDE^AN
DREW^W^Jr||19610615|M||C|1200 N
ELM
STREET^^GREENSBORO^NC^27401-
1020|GL|(919)379-1212|(919)271-
3434||S||
PATID12345001^2^M10|123456789|
987654^NC|<cr>
NK1|HYDE^KIRSTEN^G|WIFE||||||N
K^NEXT OF KIN<cr>
PV1|1|I|2000^2012^01||||004777
^JAMES^SIDNEY^J.|||SUR||||ADM|A
0|<cr>
```

The message above translates to: Patient Andrew W. Hyde, Jr., was admitted on August 18, 1988 at 11:23 a.m. by doctor Sidney J. James (#004777) for surgery (SUR). He has been assigned to room 2012, bed 01 on nursing unit 2000. The message was sent from system ADT1 at the MCM site to system LABADT, also at the MCM site, on the same date as the admission took place, but three minutes after admission.

more common way to make information available at the point-of-care. Health-care institutions also use Intranets inside the institution and Extranets between them. Charts, images and text based information are often presented on web pages. Structured data entry at the point of care is therefore feasible.

From RDE to DDC

To move from RDE to DDC it is important to throw away the old CRF model. We must adjust ourselves to capturing source data directly into a computer. The current CRF is likely to be used in only a limited number of situations - and probably have another name. Computers have been around long enough for clinical trials and health care professionals to understand and be comfortable with the issues of validation, backup and security. With careful planning, these issues present few problems.

The changes required to move forward from remote data entry are as great as, if not greater than, the changes required in moving from paper CRFs to electronic CRFs. (The eCRF is still a complete record in one place of the subject's progress through the trial). With the change to DDC, information will be more patient-based with addition information collected specific to the patient's involvement as a subject in the clinical trial. If a concept of a CRF remains, it will be of a virtual CRF. Today people in document management can now work with virtual documents in the same way. Several people write a document at the same time. Then, when it needs to be viewed as a complete document, it is "stitched" together by the document management system.

When a virtual CRF model is accepted by the clinical trials industry, it will be possible to move forward by identifying all the possible data sources and plan the best way to capture data directly. This is as much a technological challenge as a process planning challenge. Data may well exist in several machines' permanent storage or on a variety of external storage devices such as CD-ROM, diskette, DVD (the digital versatile disk, originally the digital video disk), MO (magneto optical) disk, and others. Data stored in all of these media must be captured into one system (the complexity of the task provides an interesting challenge for technology enthusiasts). Reducing the number of storage options or capturing the data over a network are alternative solutions. Capturing over a network involves using a communication standard such as HL7 or DICOM and then using the target machine to store the data in a consistent physical format.

A number of machines are being connected up to a TCP/IP (Transmission Control Protocol/Internet Protocol) communications systems so they can send data into a central, technologically homologous storage. A modern MRI machine is a good example; it can be hooked up to the Internet to send pictures wherever they are needed. For this kind of data transfer to occur there must be standards similar to the HL7 standard for communicating between imaging systems. DICOM is such a standard. It defines structures and protocols for transmission and storage so that any DICOM compatible computer can read the pictures.

If both the pharmaceutical industry and the health care industry move to DDC, the result could be a parallel effort that captures patient data into two different systems. That introduces the potential for inconsistency and an unnecessary amount of extra work - work that will be particularly noticeable to the investigator and other site staff responsible for collecting the data. To put the final piece of the puzzle in place and achieve the ideal information flow, would be ideal to first capture all the data into one system, then selectively export it to the other systems.

The future of DDC

If one assumes that a person is in the health care institution primarily as a patient and secondarily as a subject in a clinical trial, then it is logical to see the data collected first in the EPR, then exported to the sponsor as the patient-subject completes study participation. But that process raises a number of issues, not least of which are patient confidentiality, data quality issues and data coding standardization.

Confidentiality. One suggestion from the health care industry to ensure confidentiality and control is to make the data extraction from the EPR protocol driven. Clinical protocols must drive the extraction utility so that only the data required, and no more, are transferred.

Data quality. Data will be owned by the trial site - a requirement of ICH/GCP - and the quality of the information will be the site's responsibility. Any error that the sponsor discovers must first be corrected in the EPR system before being "re-extracted" to the sponsor. Here, the pharmaceutical industry's experience with front end error checking from RDE systems and the use of standard operating procedures (SOPs) to ensure data quality can be used to address quality issues in health care data systems.

Standardization. Standardization issues must be addressed to enable consistent storage and communication of the required data. The

latest version of the HL7 standard, v3.0, contains messages for clinical trials. The messages include such things as; subject registration and study completion of a patient, trial phase information, treatment schedules, sponsor information, randomization codes, subject consent information, evaluable status.

Two other standards for treatment and drug prescription information are widely used in the health care industry: the ICD9 (International Classification of Diseases, 9th revision) and SNOWMED (Systemized Nomenclature of Medicine) dictionaries. HMOs use them to assign standard codes to items that will be charged to insurance companies and also as the basis for determining the treatment/outcome for cost/benefit analysis. The pharmaceutical industry uses the World Health Organization (WHO) drug and adverse event dictionary and the new MedDRA (Medical terminology for Drug Regulatory Affairs, a resource similar to ICD10) dictionary for recording diagnosis, adverse events and reactions, and elements of product characteristics summaries.

The Health Information Management and Systems Society (HIMSS) and the Radiological Society of North America (RSNA) are working together in a joint initiative, Integrating the Healthcare Enterprise (IHE), to co-ordinate and promote standards. This is a good opportunity to promote one standard across those two disciplines.

From vision to reality

Before the pharmaceutical industry and the health care industry can capitalize on technology's potential, vendors of both EPRs and EDC systems must co-operate. Only such co-operation can ensure a standard and co-ordinated approach.

In this bright future of co-operation, one final move must be made: accepting the Internet as a safe and effective means of transferring clinical trials data. The data can flow seamlessly across the Internet from the EPR system to the sponsor. But don't look for this to happen tomorrow!

Even with the development of Web-enabled RDE systems, we're unlikely to see much innovation or trial penetration in RDE. It is more likely that DDC will be the theme for electronic data capture conferences over the next 10 years.

The change will neither be easy nor fast, but the benefits are there to be reaped, particularly for the pharmaceutical. The use of DDC, with the extraction of data from EPR systems, should produce cost and time savings. The health care industry will save because the new systems will remove much of the technological duplication now beginning to emerge in the investigator's daily routine. Clinical trials professionals will gain because data collection and error correction will be simplified during and after clinical trials.

The cost effectiveness of the DDC will depend upon the level of standardization achieved. Without a high level of standardization, each individual implementation will cost too much. One way to spread the cost of implementing a DDC program is to reuse study centers, thus spreading the cost across several studies.

Although RDE has not fulfilled new technology's potential for data capture, new technology is, and will be, an increasingly important component of data collection. To avoid duplication and technological overload for the investigators upon whom clinical trials rely so heavily, it is important to identify the technological synergies that can simplify the work of health care clinical trials professionals.

References

1. A.W., Hyde, 1998 New Technology Systems being tried in the collection of clinical trials data: a user centred comparison using HCI methods. Authors unpublished Masters Degree dissertation. (available electronically by request to the author)

Andy Hyde, *MSc (CCI) (Open), information management and technology consultant, Nycomed Amersham Imaging, Nycoveien 1-2, PO Box 4220, 0401 Oslo, Norway, +47 23185312, fax +47 23186003, e-mail: ahy@nycomed.com.*