MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press
© 1998 IMIA. All rights reserved

A Quantitative Perspective on the Virtual Patient Record (VPR) and its Realization

Jochen R. Möhr

School of Health Information Science
University of Victoria, Victoria, B.C., Canada

Abstract

The Virtual Patient Record (VPR), the union of all collections of health relevant data that accumulates over a person's lifetime in any institution that person has contact with, is a technical possibility in the age of networked computers[1]. This paper investigates, under what conditions the VPR may be useful and manageable. Quantitative considerations of the VPR for a population of 10 M people form the basis. Dynamic aspects of updating and decay in value for treatment decisions are also taken into account. The role of centralized and distributed data stores are compared and the need for indexing is identified.

Keywords:

Virtual patient record; Electronic Patient Record; Data warehousing; Record archiving; Indexing;

Introduction

The vision of the Virtual Patient Record (VPR) [1] has been with us for a while. The assumption was always that it would be useful and usable. But is it really so, and if so, how? The following paper is a contribution to the discussion of these issues.

A Hypothetical Model of the VPR

In the following a number of assumptions are made on the qualitative and quantitative characteristics of an individual VPR and those of an aggregate of such records that serves a population of 10 million people. For the purpose of this discussion the virtual record is defined as that set of data that has been completed at a given point in time. This would include all completed "contacts", e.g., the last x-ray investigation complete with report, the last consultation report, but would exclude a set of lab investigations that has not been confirmed yet. In a sense the perspective of a user that accesses the VPR like one would access the conventional patient record, a file that includes all completed investigations at that point in time, is adopted.

Qualitative Data Composition

Qualitatively one can assume that the VPR will be a multimedial record including much of the contents of the conventional record plus electronic components. This discussion covers only the electronic data component which will increase over time and may eventually dominate. This will consist of formatted data stored in relational data base fashion. Other objects will include free text, either dictated or synthesized from formatted data, or from the automated interpretation of biosignal tracings and images. The extraction of text from paper documents through OCR technology may add to the volume of electronically stored text over time, as well as to the volume of formatted data. Biosignal tracings and images will be incorporated as such and may include sounds, document images, etc.

Quantitative Data Composition

The quantitative estimates used here rely on analyses from Europe and North America (compare, e.g., [2-4]), and unpublished observations of the author. The numbers are considerate estimates attempting to characterize a complex situation and do a multitude of factors justice. They should be thought of as averages of heavily skewed frequency distributions with a modus below the average. We also distinguish three care levels: normal health, acute care, and maximal or chronic care. For our discussion, we assume the characteristics given in Table 1 for the corresponding populations. The institutions include physicians, dentists, pharmacies, labs, and potentially schools, employers, etc., in the population with normal health, and hospital departments, such as laboratory, imaging, etc., in the other care categories. It is assumed that each of these institutions store 2KB of demographic data on each person. The storage of 10, 200 and 300 KB of other formatted data is assumed in 60 50 and 80% of institutions for the three care levels respectively. The storage of a text data volume of 25, 100 and 300KB of text is assumed in each institution for the three care levels. The averages given for images and biosignal tracings, on the other hand are assumed for all institutions. Overall, our assumptions are conservative.

Table 1 - Data Characteristics of Persons in Normal Health, Acute Care and Maximal/Chronic care

	Normal Health	Acute Care	Max./Chron. Care
Age	35	55	65
Number of Institutions	10	25	40
Formatted demogr.	2KB	2KB	2KB
Formatted other	10KB	200KB	300KB
Text Data	25KB	100KB	300KB
Images & Tracings	12 images 0 tracings	40 images 20 tracings	100 images 100 tracings

Table 2 - Average data volume per person in normal health, acute care and maximal/chronic care

	Normal Health		Acute Care		Maximal/ Chronic Care	
	MB	%	MB	%	MB	%
Formatted Data	0.08	3	2.55	17	9.68	19
Text	0.25	9	2.50	17	12.00	23
Images	2.40	88	10.00	66	30.00	58
Total	2.73	100	15.05	100	51.68	100

Our assumptions translate into the total data volumes for each care level shown in table 2. For individuals in normal health, an average data volume of 2.73 MB results, almost 90% of which are image data. In acute care individuals, the total volume increases about five fold, and the relative contribution of formatted and text data increases to almost 35%. This trend continues in maximal/chronic care, the total volume being almost 20 times that of normal health, and formatted and text data contributing just above 40%. Because of the skewed distributions characterizing such data, one has to assume that the extremes lie in the order of 10 to 100 fold these averages. This volume is easily accommodated by networked computers. It may, however, be a challenge for patient carried devices. the

Next we upscale these data for a population of 10 million, assuming that we are converting to the VPR at a given point in time. The situation will then represent a snap shot of

morbidity pattern at that time, for which we assume that 80% of the population are characterized by data patterns consistent with normal health, and 15 % and 5% respectively by those of acute care and maximal/chronic care. The resulting data volume is 70 TB (Tera Bytes), of which about one third is contributed by each of the distinguished care categories. Eventually, as the system reaches maturity, more people will have experienced bouts of acute care or maximal care. The population will have aged and more people will be treated chronically. The number of institutions that store data electronically and the amount of data per person will also likely have increased. For our discussion we look at three stages: the "initial stage" with the distribution of care categories assumed so far (80:15:5%), an "intermediate stage", with a distribution of 60:30:10%, and a "late stage" with 40:30:30% (see Table 4). The late stage is included in order to explore extremes. Even under these extreme assumptions the total data volume increases only by a factor of 2.5 to 175 TB. Again, this data volume is no challenge for networked computers, and could even be accommodated in one dedicated institution.

Table 3 - Average data volume in population of 10 Million in Tera Bytes (TB)

	Normal Health	Acute Care	Maximal/ Chronic Care	Total
Percent of population	80%	15%	5%	
Formatted Data	0.64	3.83	4.84	9.31
Text	2.00	3.75	6.00	11.75
Images	19.20	15.00	15.00	49.20
Total TB	21.84	22.58	25.84	70.26
% of Total	31%	32%	37%	100%

Table 4 - Total data volume with increasing maturity of electronic patient record

Stage	Initial	Intermediate	Late
Care Categories:			
Normal Health	80%	60%	40%
Acute Care	15%	30%	30%
Max./Chron. Care	5%	10%	30%
Total Data (TB)	70	113	175

Dynamic Considerations

VPR Updates

Let us next take a look at the dynamic aspects of updating and retrieving data from this hypothetical virtual medical record of a population of ten million people. We equate "update" with "contact". This leads to a low estimate, and the number of updates of the data base in a single transaction oriented information system contributing to the VPR is many orders of magnitude higher. But the estimate is consistent with our external user view of the VPR.

In Table 5, we assume five contacts/updates per year for people in normal health, ten for those in acute care, and fifty for those in maximal/chronic care. If these estimates are not spread as widely as one might expect, it is due to the consideration that

Table 5 - Updates in Population of 10 Million

		Initial	Inter-mediate	Late
Care Level	Contacts	Updates per year		
Health	5	$40x10^6$	$30x10^6$	$20,x10^6$
Acute	10	$15x10^6$	$30x10^6$	$30x10^6$
Max./ Chron.	50	$25x10^6$	$50x10^6$	$150x10^6$
Total/yr.		$80x10^6$	$110x10^6$	$200x10^6$

contacts in normal health include visits to the optometrist, etc., while maximal care takes place in intensive care institutions, where updates can be accomplished with major chunks of data for comprehensive treatment episodes. The same category includes chronic care with long episodes of relatively low data activity. Accordingly, we arrive at 80 to 200 million updates per year, an unimpressive figure.

Decay of Data Value

The value of data determines whether it is worth to store, access and retrieve it. The value is specific to an information purpose and decreases with time. For treatment decisions, the value of medical data decays rather rapidly. This is obvious if one considers the mechanisms that govern many data, such as serum electrolytes, blood pressure, the demonstration of an ulcer on a radiograph. Rapid decay is also demonstrable from access rates to archival data [5]. Since patient care purposes are very diverse, it is not easily predictable what data might be of value in the future. Although few actual measurements exist, one can probably assume that the average half life of medical data is less than three months for treatment decisions. For other purposes, particularly research, the half life is much longer. Due to the low basic data quality, in particular the lack of completeness of routine data, however, the value of the bulk of medical data for research is extremely low to begin with.

Table 6 gives an overview of the decay we can expect under the conditions considered so far. For the initial stage of the virtual record it is assumed that the data collection is relatively young. For thirty, forty and thirty percent of the data an average age of 5, 2 and 1 years are assumed respectively. A half life of three months leads to a residual mass of data that is valuable for treatment decisions in the order of 4.6 TB, or 6.6% of the total of 70 TB. For the intermediate and late stages, an increasing age of the data is taken into consideration. As a consequence, residual volume of valuable data decreases to 0.24%, and then to less than 0.01%! T

For our purposes the absolute magnitude is considerably less important than the insight into the dynamics: while data accumulate gradually to an appreciable but manageable level, their value decreases drastically. This may raise the question of whether it is worth storing the data in the first place over such time. But it also points to a potentially more serious problem: finding relevant data will amount to searching for the proverbial needle in the haystack, unless proper precautions are taken.

Table 6 - Decay of Data Value (Assumed Data Half Life 3months)

Stage:			
Initial	**Total volume 70TB**		
	Av. Age	%	Useful MB
	5	30	66.7572
	2	40	273,437.5000
	1	30	4,375,000.0000
Total MB			4,648,504.26
% useful			**6.6407**
Intermediate	**Total volume 113TB**		
	Av. Age	%	Useful MB
	10	50	0.0001
	5	30	66.7572
	2	20	273,437.5000
Total MB			273,504.26
% useful			**0.2420**
Late	**Total volume 175TB**		
	Av. Age	%	Useful MB
	30	60	0.0000
	15	30	0.0000
	3	10	17,089.8438
Total MB			17,089.84
% useful			**0.0098**

Potential Solutions

If the total volume of data of the VPR is less of a problem than the relative scarcity of useful data contained in it, our aim should be to:

1. increase or preserve the value of the data, or
2. improve the accessibility of data that are of value.

Since the purposes in patient care are as varied and diverse as health problems it is difficult to select or abstract data so that their value for patient care is preserved, let alone increased. But it would make sense to increase the value of data for other purposes, such as research. This could be accomplished by increasing the completeness and accuracy of data, e.g. by archiving original recorded data, such as lab measurements, biosignals, and images. This, however, runs contrary to the second solution principle, that of improving data accessibility.

In treatment support, data are usually accessed using guidance from simple structures, such as time and source within a person's record. Most research requires access across persons. This could be improved by the selection of summary data through abstraction or indexes. Both approaches are closely related since abstracted data could serve as index. Since there is a limited capacity for human abstracting, automated means for high quality abstracting and indexing are crucial. In the final section

we therefore investigate the potential contributions of central versus distributed data stores, and human versus automated indexing to a solution.

Central Stores

Central stores of data are a fashion revived in the Nineties under the label of data warehousing [6]. It is defined as the enterprise wide subject-oriented, integrated collection of time invariant, and non-volatile data resources [6]. Attractive features of this approach are the potential for a central, well organized data repository that is administered consistently and available to anyone with a legitimate need. The legitimization, data security, and confidentiality could be enforced uniformly, and automated indexing applied consistently, e.g. at discrete update intervals, to the entire collection.

Even if one takes into consideration that the central storage requirements will be a duplication, or more likely a manifold of the distributed storage requirements, in order to accommodate backup, security copies, and software required for the management of the data collection, the storage capacity is not an issue within current technical capabilities. This means that the warehouse approach could be considered provided that access to the data can be accomplished, and that non-technical issues do not prevail.

However, ethical and political issues are likely to be important detractors from the idea of a centrally administered store of all detailed medical data on every person. It is also likely that the reliable realization of the advantageous features of the data warehouse would lead to a massive administration. Many forces, e.g., economical, administrative, compatibility issues, etc., could combine to result in a technologically sub-optimal solution in such an administration. There is also always a danger that data of such a central administration are usurped for purposes that are not legitimate, let alone ethical. But even apart from the danger of technical inappropriateness and uncertain long term reliability of such an apparatus, it is questionable whether such a massive effort makes economic sense, given the low data value for treatment decisions and research.

Distributed Stores

Distributed stores already exist in the form of the institutions contributing to the VPR. Much administrative overhead and multiplication of storage requirements could be avoided if it was possible to use these existing stores as such as the VPR. This would require that data security, protection of privacy, and access to needed information be provided reliably on such basis. Data security is not a new issue for participating institutions and would have to be assured regardless. Protection of privacy is - if not a new - at least an augmented issue since access to data would have to be provided outside the institution proper, and such aspects as ascertaining legitimization of requests and requesters would take on a different magnitude. In addition, the privacy issue becomes more complex because not only is the privacy of the patient at stake but also the provider's. Also, the linking of classes of users to privileges pertaining to classes of data does not really suffice and should be complemented by access privileges specific to a purpose of an information request. This is where guarding of privacy merges with provision of access since one mechanism is essentially the complement of the other; if we are able to provide access for legitimate purposes, we are also able to deny access for illegitimate ones.

The provision of access itself may however turn out to be the most elusive problem. Differences in information models pose greater problems than differences in data formats. Data exchange between heterogeneous systems through standards like HL7 is a great accomplishment for objects defined at the field level, as in relational data bases. It does not go far enough for complex data like texts, biosignal tracings or images. It does not cover the purpose of a data request. Abstracting (or indexing) mechanisms would thus be required to mediate and therefore become crucial components of the VPR and shall be examined next.

Human Indexing

Indexing, the identification of important aspects of the medical record by mapping to a classification such as ICD is standard in hospitals. It is complemented by a routine activity of many family physicians, not usually considered abstracting: the extraction of "important" data from external documents such as consultation reports for inclusion in the record. This leads to information concentration of through selection by highly skilled intelligence. It is a model for a condensed summary record that could serve to provide pointers to more complete stores of original data. It could provide the basis for supporting family physicians in a gatekeeper role with control over contents of, and perhaps even access to, the VPR. Since this might exceed the capacity of the family physician, other professionals, e.g., health informaticians could be assigned this role.

Automated Indexing

Automated indexing is the analogous extraction of meaning from medical documents. It is accomplished by mechanisms such as the automated classification of biosignals or images, by free text analysis, or the mapping of free text to defined classifications, and also by the synthesis of features recorded in a relational database into reports or abstracts. The problem is not one of basic feasibility but rather of the large diversity of approaches that can, or have to, be applied to the diversity of medical data. We are currently exploring the potential contribution of statistical approaches to document indexing in this context [7] because it could be applied uniformly to free text – regardless of whether it is composed by humans, e.g., through dictation or generated by automated interpretation of human observations or recorded biosignals and images. It could therefore serve as a standard approach to comprehensively cover a large part and important aspects of the VPR. It could thus complement the human abstracting.

Conclusions

VPRs as data collections are feasible. But that does not make them useful. The ability to access specific data is more of an issue than storage capacity. Non-technical issues are of greater concern than technical ones. The appropriate combination of central and distributed stores with automated and human indexing, and human gate keepers may provide a solution. A central

store could serve as archive for research of all anonymised original measurements, e.g., data from labs, electro-physiology and imaging. This data could serve as backup for patient care in the rare instances it is needed. In such cases, the data could be linked to a patient identity contingent on a protocol, e.g., contingent on patient care need and patient consent. The central stores could use automated abstracting and indexing mechanisms to provide pointers to the contents of the data. These pointers could improve the value of the stored data for research and management. Routine patient care could use shorter term stores with data abstracted by human gate keepers, such as family physicians or health information professionals.

Acknowledgments

This research was supported in part by HEALNet Health Informatics and a grant from the Province of British Columbia. I appreciate the support of Chris Anglin in proof reading the text, and of Sokvider Kaur in complying with the format specifications. Much inspiration resulted from discussions with students and HEALNet colleagues, in particular Yuri Kagolovsky, David Freese, Mike Miller and Toby Walrod.

References

[1] Forslund DW, Phillips RL, Kilman DG, Cook JL. Experiences with a Distributed Virtual Patient Record System. *JAMIA Symposium Supplement* 1996 483-7.

[2] Möhr JR, Haehn KD ed. *Verdenstudie - Strukturanalyse Allgemeinmedizinischer Praxen.* Koeln: Deutscher Aerzteverlag, 1977.

[3] Wiederhold G. *An Analysis of Automated Ambulatory Medical Record Systems.* San Francisco: Univ. of California, 1975.

[4] Reichertz PL, Möhr JR, Schwarz B, Schlatter A, von Gärtner-Holthoff G. Evaluation of a Field Test of Computers for the Doctor's Office. *Meth Inform Med* 1989: 18 61-70.

[5] Dujat C, Haux R, Schmucker P, Winter A. Digital Optical Archiving of Medical Records in Hospital Information Systems - A Practical Approach Towards the Computer-based Patient Record? *Meth Inform Med* 1995: 34 489-97.

[6] Inmon W, Hackathorn R. *Building the Data Ware-house.* New York: Wiley-QED, 1992.

[7] Kagolovsky Y, Miller M, Möhr JR. Statistical Concept Representation for Indexing of Clinical Narratives. In Fisher, P ed. *COACH Conference 22 Scientific Program Proceedings.* Edmonton: HC&CC 1997 118-126.

Address for Correspondence

Jochen R. Moehr
School of Health Information Science
University of Victoria
Victoria, B.C.
Canada V8W 3P5
jmoehr@hsd.uvic.ca
http://www.hsd.uvic.ca/HIS/his.htm