BigMouth: a multi-institutional dental data repository

Muhammad F Walji,[1] Elsbeth Kalenderian,[2] Paul C Stark,[3] Joel M White,[4] Krishna K Kookal,[5] Dat Phan,[6] Duong Tran,[5] Elmer V Bernstam,[6,7] Rachel Ramoni[8]

For numbered affiliations see end of article.

Correspondence to
Dr Muhammad F Walji, Department of Diagnostic and Biomedical Sciences, The University of Texas Health Science Center at Houston, School of Dentistry, 7500 Cambridge Street, Suite 4.184, Houston, TX 77054, USA; Muhammad.F.Walji@uth.tmc.edu

MFW and EK contributed equally.

ABSTRACT
Few oral health databases are available for research and the advancement of evidence-based dentistry. In this work we developed a centralized data repository derived from electronic health records (EHRs) at four dental schools participating in the Consortium of Oral Health Research and Informatics. A multi-stakeholder committee developed a data governance framework that encouraged data sharing while allowing control of contributed data. We adopted the i2b2 data warehousing platform and mapped data from each institution to a common reference terminology. We realized that dental EHRs urgently need to adopt common terminologies. While all used the same treatment code set, only three of the four sites used a common diagnostic terminology, and there were wide discrepancies in how medical and dental histories were documented. BigMouth was successfully launched in August 2012 with data on 1.1 million patients, and made available to users at the contributing institutions.

INTRODUCTION

The purpose of this work was to develop a data repository called BigMouth, which is derived from electronic health records (EHRs) at four dental schools at the University of Texas Health Science Center at Houston (UTHealth), Tufts University, the University of California San Francisco (UCSF), and Harvard University. We aimed to establish the technical foundation, develop a data governance framework, and address areas that need to evolve in order to facilitate the use of the repository for clinical and quality improvement research.

Over 160 000 active dental practitioners provide care to more than 83 million patients each year in the USA,[1] yet there is currently little understanding of the effectiveness of common dental treatments or the relationship between oral and general health. As is the case in medicine, trials are costly. However, secondary uses of data already stored in dental EHRs have great potential to improve the data-driven knowledge base in dentistry[2] and answer basic questions such as 'how long do tooth-colored fillings last?' and 'how often do patients with diabetes receive the recommended periodontal screenings?' Linking data from dental EHRs with medical EHRs may also clarify the relationship between oral and general health.[3]

Dentists have a long history of using computers for clinical tasks.[4] The overwhelming majority of dental schools in the USA have already adopted EHRs.[5][6] In 2013, 73.8% of solo practitioners and 78.7% of dentists in group practices participating in dental practice-based research networks (DPBRNs)

used computers to manage clinical information.[7] Several dental EHR systems, including Dentrix, EagleSoft, and SoftDent, are used in private practice. However, the majority of US dental schools use the same commercial EHR platform: axiUm, which was acquired by Henry Schein, the makers of Dentrix, in 2012.

CASE DESCRIPTION

The dental schools at UTHealth, UCSF, Harvard, and Tufts belong to the Consortium of Oral Health Research and Informatics (COHRI) which was formed in 2007 by users of the axiUm dental EHR.[8] COHRI institutions share best practices and develop standardized data collection tools with the aim of using data already collected in the EHR to improve oral health research, education, and treatment. One of the early tasks of the consortium was to address the difficulty researchers at the individual institutions experienced in querying the EHR at their local site. The development of a centralized inter-institutional data repository would allow faculty, residents, and students access to a large dataset to conduct retrospective studies. There was also a need to develop a data governance approach that encouraged dental institutions participating in COHRI to contribute data.

METHODS

Establishment of the BigMouth Dental Data Repository

While no site had developed a dental repository at the time, two of the participating sites had already implemented the Informatics for Integrating Biology and the Bedside (i2b2) data warehousing platform[9] as part of their respective Clinical Translational Science Award (CTSA)-related initiatives. BigMouth differed from a typical i2b2 implementation[10–16] because the purpose was to develop a centralized repository for four institutions. An advantage of i2b2 was the availability of a web-based client that would allow end users from each of the institutions to securely access, explore, and query the repository. This project was funded by the National Library of Medicine.

Data governance

A multi-stakeholder committee made up of representatives of the four dental schools and COHRI developed five principles for data governance. The following principles were framed to support the goals of site enrollment, data improvement, and clinical translation.

Adhere to privacy and security requirements

BigMouth contained a 'limited dataset' where data are de-identified with the exception of dates (eg, date of birth, visit dates) and zip codes. According to the requirements of the Health Insurance Portability and Accountability Act (HIPAA) privacy rule, each participating institution must execute a Data Use Agreement (DUA) to allow the use of a limited dataset for research purposes.[17] No further authorizations from patients are required for a limited dataset. In addition, each contributing institution needs IRB approval.

Access provided to those institutions that contribute

For any organization to access the data, they must also contribute to the repository. Users from each site may use the i2b2 Web Workbench to query their own data as well as a combined dataset from all sites. A user cannot associate the data with the contributing institution when querying the combined data.

Each source site retains control of contributed data

Each contributing institution controls the data it contributes and has the right to remove data. However, data cannot be removed retrospectively from approved research projects where data have already been shared or committed to be shared. New projects will not have access to data from the school that decided to stop sharing data.

Project review committee approval required for use of data for specific research projects

A project review committee with a representative from each contributing school as well as other members of COHRI will coordinate research projects among the sites. Researchers submit proposals to the project review committee.

Continuously assess and improve the quality of data

A data quality checklist will be executed each time data are loaded into the repository. This checklist identifies missing data and determines the count of patients retrieved for specific queries (eg, number of patients with a medical history) and compares it against the previous data load. Any discrepancies such as a loss in patient count are further investigated.

Development of the BigMouth Dental Data Repository

A two-phased development approach was used to explore, extract, load, and map data from the respective sites into the shared repository. Although the sites used the same EHR, they were implemented and coded differently. While each site collected unstructured and structured data in their dental EHRs, we loaded only structured data into BigMouth (see table 1). In phase 1, scripts were developed to extract these data from each site's EHR and generate a file suitable for loading into i2b2. The EHR vendor provided a data dictionary and assisted with understanding how data were stored in the system.

In phase 2 we mapped the unique concepts from each local terminology to build a reference terminology, called the COHRI terminology. At the database level, BigMouth contained the source data from each site, thereby maintaining provenance. The COHRI terminology allowed users to query data across sites without being able to identify the specific institutions through the use of access controls at the user interface level. In instances where a standardized terminology was used, the mapping was relatively straightforward, for example, the Current Dental Terminology (CDT) was used to map all procedure codes. However, even in this context, each site had made customizations to the CDT codes, such as adding granularity to allow for documentation of a procedure's sub-steps. This is useful in the teaching setting where a single procedure may take several visits. The root CDT code for the procedure would be assigned at the visit during which the procedure was completed. In the context of procedures, locally generated codes were ignored, and so BigMouth captured only the completed procedure. The EZCodes Dental Diagnostic Terminology was used to map the diagnoses.[19] Adults have 32 adult teeth and children have 20 deciduous or primary teeth: all four institutions used the Universal Tooth Numbering System to refer to each tooth.[20] In the dental clinic, teeth that are missing, have previously been treated, or have disease/injury are charted on the odontogram, a visual representation of the teeth. Three of the four institutions used the axiUm EHR's default odontogram. Periodontal data had even greater consistency among the sites.

The medical and dental histories were extensive and varied greatly among sites. Across the sites, a total of 975 questions were used to capture the medical history and 284 questions to

Table 1 Data captured in BigMouth

Type of data	Description/example	Use of common terminology or standard forms
Demographics	Age, race, ethnicity, sex	One site used the OMB/NIH definition for race and ethnicity,[18] others had unique definitions.
Diagnoses	Dental diagnoses such as root caries, generalized moderate chronic periodontitis	Three sites used the EZCodes Dental Diagnostic Terminology.[19] One site did not document diagnoses in a structured format.
Medical history	Patient's medical history including vital signs	Great variance in the number and type of data collected among sites. One site used the COHRI standardized medical history form.
Dental history	Patient's dental history of pain, periodontal problems, or previous oral surgery	Great variance in the number and type of data collected among sites. One site used the COHRI standardized dental history form.
Procedures	Dental procedures or treatments conducted such as biopsy of oral tissue, removal of impacted tooth	All sites used the CDT codes. However, each site had made local customizations.
Odontogram (tooth chart)	Observations relating to the teeth and periodontium (gingiva, oral mucosa, and bone)	The Universal Tooth Numbering System was used by all four sites. Three of the sites also used the default data collection forms that came installed in the EHR to describe existing conditions and materials for each tooth.
Periodontal chart	Bleeding on probing, probing depth, and gingival recession	All four sites collected the same core measures in the same way. Some sites collected additional measures.
Treating provider	Dental student, resident, hygienist, faculty dentist	While dental student and faculty dentist roles were the same across all four sites, classification of other personnel such as hygienists, assistants, and radiology technicians varied.

CDT, Current Dental Terminology; COHRI, Consortium of Oral Health Research and Informatics; OMB/NIH, Office of Management and Budget/National Institutes of Health.

Table 2 Demographic characteristics, oral health status, and selected procedures of patients in the clinics of four dental schools in the BigMouth Dental Data Repository database between January 1, 2010 and December 31, 2011

	School of dentistry			
	School 1	School 2	School 3	School 4
Demographics	N=15 219	N=34 126	N=34 318	N=13 927
Mean age (SD)	48 (17.0)	47 (17.8)	50 (23.2)	45 (17.5)
Sex (%)				
Male	42.4	46.1	45.3	39.2
Female	55.7	53.9	53.3	55.9
Others/don't know	1.9	0.0	1.4	4.9
Diagnosis	N=6227	N/A	N=10 451	N=3775
Defective restoration (%)				
Open margin	4.6		5.6	1.7
Removable prosthodontics (%)				
Partially edentulous maxilla	2.2		1.7	4.0
Forms	N=11 171	N=24 715	N=20 942	N=3588
Dental history (%)				
Sensitive to cold, hot, sweet, or pressure	19.0	16.1	37.7	12.9
Medical history (%)				
Hypertension	9.8	13.4	14.4	21.5
Oral health status	N=15 219	N=34 126	N=34 318	N=13 927
Missing teeth				
Mean number of missing teeth (SD)	4.1 (5.7)	5.8 (7.5)	5.2 (6.6)	4.1 (6.7)
	N=5698	N=8641	N=8128	N=2434
Dental caries (%)				
Dental caries	75.5	72.2	71.1	84.0
	N=913	N=2671	N=4918	N=5913
Periodontitis (%)				
Periodontitis	54.9	78.5	87.8	71.3
Procedures	N=14 526	N=30 732	N=32 163	N=13 594
Diagnostic X-ray (%)				
Intraoral X-ray-complete series	20.7	21.0	20.6	19.0
Preventive procedure (%)				
Prophylaxis	42.5	39.8	34.5	17.5
Therapeutic procedure (%)				
Extraction, erupted tooth, or exposed root	7.8	23.3	3.1	21.8

capture the dental history. Only nine concepts were found to be exactly the same across the four sites in the medical history form: a patient's history of (1) diabetes, (2) alcohol consumption, (3) pregnancy, (4) rheumatic fever, (5) myocardial infarction, (6) hypertension, (7) seizure, (8) asthma, and (9) allergy to local anesthetic. Only seven dental history concepts were captured at all of the four sites: (1) sensitivity to cold, hot, sweet, or pressure, (2) difficulty or pain upon chewing, talking, or using the jaw, (3) clenching or bruxing teeth, (4) having partial or complete dentures, (5) history of periodontal treatment or surgery, (6) history of root canal treatment, and (7) history of braces or orthodontic work. The lack of standardization or the use of common terminologies posed the biggest challenge to mapping.

Access to i2b2 web client for end users

In order to allow end users (faculty, residents, students) to access the i2b2 web client, we used the InCommon federation (incommon.org) which provides a secure and trusted identity management system for member universities and other research institutions. InCommon leverages Shibboleth, a standards-based open source system, which allows authorized users to use their existing institutional login and password or use a third party

(Protect Networks) to access BigMouth. The use of InCommon also enabled us to set group authorization (see figure 1).

After authenticating via the i2b2 web client, users can either explore data from their own site (local terminology) or the data from all four sites (COHRI terminology). The data for each site were modeled so that a user familiar with the axiUm EHR would be able to easily navigate through the i2b2 representation. We included all concepts that were collected at each site in the same hierarchy and order that was present in the source EHR. This was especially useful for the custom forms, which each site used to represent the medical and dental history sections. This allows users to quickly find concepts using the same mental model that they are used to when documenting in the EHR.

Table 2 demonstrates the similarities and differences between the sites on a cross-section of data contained in BigMouth. Patients receiving care at the sites have a high prevalence of both dental caries and periodontitis. Use of diagnostic X-rays, preventative and therapeutic procedures differ among the sites. These differences represent opportunities to learn from one another, thereby identifying best practices for data collection, if discrepancies are due to poor documentation, or for evidence-based care.
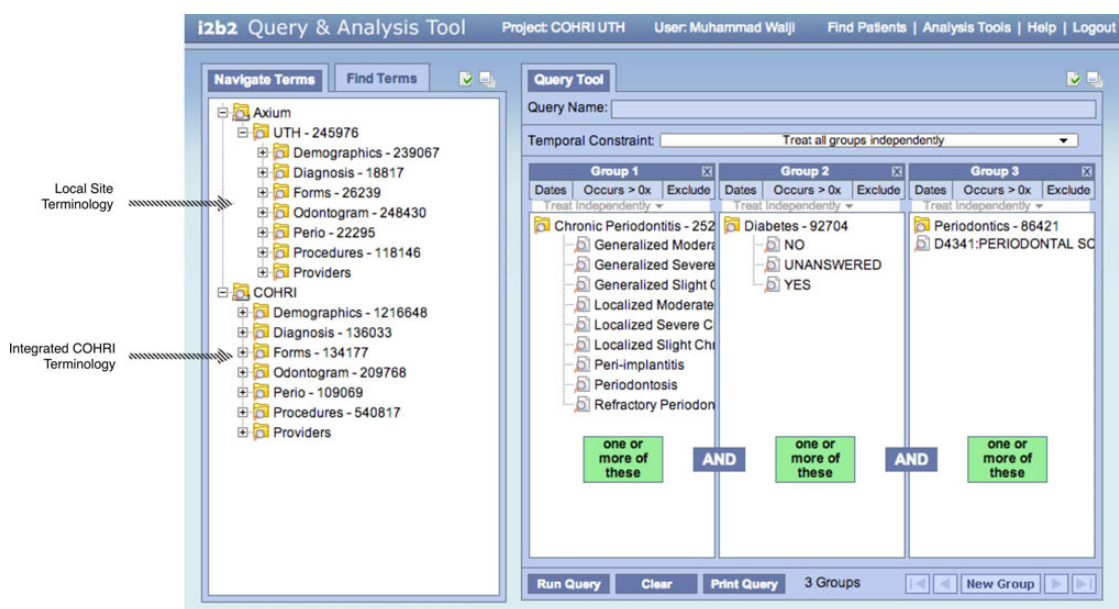
Figure 1 i2b2 web client view of BigMouth. Users are restricted to exploring and querying their local site data or integrated data from all four sites (COHRI, Consortium of Oral Health Research and Informatics).

Launch of BigMouth

In August 2012, we launched the BigMouth Dental Data Repository with data on 1.1 million patients and made it available to all faculty, residents, and students in the four dental schools. BigMouth data are refreshed every 3 months. In addition to accessing the i2b2 web client, users have also submitted research requests to the Project Review Committee.

DISCUSSION

We were successful in creating BigMouth, a centralized data repository that contains dental EHR data from four dental schools. In so doing, we have established the largest multi-institutional dental clinical data repository. This resource is now available to end users at these institutions. In anticipation of the expansion of BigMouth to other COHRI institutions, we developed a data governance framework that encourages institutions to contribute a limited dataset while at the same time maintaining control of how these data are used for research purposes.

This significant, but first, step revealed some of the work that should be done to maximize the efficient and valid secondary use of data collected through dental EHRs. Our work, for example, highlights a pressing need for the development and adoption of data standards in dentistry. One of the most obvious gaps for the dental profession has been the lack of a standardized dental diagnostic terminology, which of course has implications for clinical care as well as the secondary use of data.[21] In response to this need, our research group has made significant progress in establishing the Dental Diagnostic System (formerly EZCodes Dental Diagnostic Terminology) as a standard both within and beyond the COHRI sites.[19] Spurred by the need for standardization so that cross-institutional clinical and quality improvement research can be conducted, COHRI has also formed workgroups to standardize the medical and dental history forms for adults and children, as well as caries ('cavities') risk assessment. However, while dental institutions that contribute data to BigMouth are encouraged to make use of these standard forms and terminologies, they are not required to do so. Expanding the data governance framework to mandate the use of these standardized tools would contribute to improved data quality, if accompanied by implementation that supports valid data entry. Outside of the strictly dental context, Meaningful Use incentives, which several dental schools are pursuing, have encouraged the adoption of standards and structured data collection, for example, patients' smoking status and vital signs.[22] We have found that BigMouth itself can be part of a learning lifecycle in which feedback to sites can improve the quality of data collected. For example, we discovered that one site had more missing race data than the other three sites. At this site, this information was collected by front desk personnel, who reported being uncomfortable collecting such data. As a result, these items were included as part of the medical history forms to be collected by the provider.

Another issue to be addressed as the secondary use of dental clinical data matures is modeling the high dimensionality data that are an essential part of the routine care of mouths that have up to 32 teeth and generate an order of magnitude more units of evaluation. For instance, in a full periodontal exam, a fundamental component of a complete oral health examination, bleeding on probing, recession, and periodontal pocket depth data are collected at six sites per tooth; thus there are 192 observations for each of these measures. Rather than represent each periodontal clinical observation, the periodontal data as a whole were captured in the i2b2 blob format. For example, if a patient had an initial periodontal examination where probing depth was documented, the i2b2 blob field would store the 192 data points. The advantage of this process was that it simplified the data extraction and load procedure. The disadvantage of the blob approach is that users cannot directly query for patients that meet specific periodontal measurement criteria.

Oral health is perceived as separable from general health in the USA.[23 24] Unfortunately, there is often a technical and policy-based firewall between medical and dental data, with no common methods of communicating efficiently across the divide. As noted by Powell and Din, 'The essential core improvement to bring medicine and dentistry closer together is the integration of medical and dental care and data. Currently, many medical records and data exist separate and distinct from dental records and data for the same patient.'[25] The

circumstances at the Harvard School of Dental Medicine exemplify the divide: Harvard has affiliated hospitals (eg, Brigham and Women's Hospital, Boston Children's Hospital), but Harvard itself provides no medical care. Even in the context of clinical care, there are no channels through which to push data to or pull data from the medical EHRs. Thus, in practice the little information that is exchanged between the medical and dental EHRs in this context is typically conveyed through letters or telephone calls. Information exchanged in this way can be included in the record only as a PDF or image or as free-text note entries. In 2009, the American Dental Association announced an agreement with HL7 (Health Level 7) to enhance the coordination of patient care between medical and dental practices using a dental extension to the Continuity of Care Document (CCD).[26][27] We look forward to well-structured dental CCDs being regularly exchanged. From the secondary use perspective, the interim recourse is statistical approaches to matching data likely to belong to the same individual across datasets, assuming one can obtain the human subjects approval to access datasets with sufficient information to perform such matching.

Much has already been accomplished with the successful limited launch of BigMouth, but challenges remain to creating the infrastructure and processes to support routine inter- and intra-professional secondary clinical data use. In the near term, we look forward to expanding the types of data, for example, medications, and the number of institutions that contribute data to BigMouth.

Author affiliations

[1]Department of Diagnostic and Biomedical Sciences Office of Technology Services and Informatics, School of Dentistry The University of Texas Health Science Center at Houston

[2]Oral Health Policy and Epidemiology Harvard School of Dental Medicine, Boston, Massachusetts, USA

[3]Department of Public Health and Community Service, Tufts University School of Dental Medicine, Boston, Massachusetts, USA

[4]Department of Preventive and Restorative Dental Sciences, School of Dentistry, University of California, San Francisco, CA, USA

[5]Office of Technology Services and Informatics, School of Dentistry The University of Texas Health Science Center at Houston, Houston, Texas, USA

[6]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

[7]Division of General Internal Medicine, Department of Internal Medicine, Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, USA

[8]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

REFERENCES

1 ADEA. Distribution of Dentists in the U.S. by Region and State 2009. http://www.adea.org/publications/library/otherresources/Pages/2007DistributionofDentistsintheUnitedStatesbyRegionandState%28December2009%29.aspx

2 Song M, Liu K, Abromitis R, et al. Reusing electronic patient data for dental clinical research: a review of current status. J Dent 2013;41:1148–63.

3 Boland MR, Hripcsak G, Albers DJ, et al. Discovering medical conditions associated with periodontitis using linked electronic health records. J Clin Periodontol 2013;40:474–82.

4 Schleyer TK, Thyvalikakath TP, Spallek H, et al. Clinical computing in general dentistry. J Am Med Inform Assoc 2006;13:344–52.

5 Atkinson JC, Zeller GG, Shah C. Electronic patient records for dental school clinics: more than paperless systems. J Dent Educ 2002;66:634–42.

6 Powell J, Buchan I. Electronic health records should support clinical research. J Med Internet Res 2005;7:e4.

7 Schleyer T, Song M, Gilbert GH, et al. Electronic dental record use and clinical information management patterns among practitioner-investigators in The Dental Practice-Based Research Network. J Am Dent Assoc 2013;144:49–58.

8 Stark PC, Kalenderian E, White JM, et al. Consortium for oral health-related informatics: improving dental research, education, and treatment. J Dent Educ 2010;74:1051–65.

9 Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17:124–30.

10 Marsolo K, Corsmo J, Barnes MG, et al. Challenges in creating an opt-in biobank with a registrar-based consent process and a commercial EHR. J Am Med Inform Assoc 2012;19:1115–18.

11 Segagni D, Tibollo V, Dagliati A, et al. CARDIO-i2b2: integrating arrhythmogenic disease data in i2b2. Stud Health Technol Inform 2012;180:1126–8.

12 Natter MD, Quan J, Ortiz DM, et al. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J Am Med Inform Assoc 2013;20:172–9.

13 Armstrong AW, Reddy SB, Garg A. Novel approach to utilizing electronic health records for dermatologic research: developing a multi-institutional federated data network for clinical and translational research in psoriasis and psoriatic arthritis. Dermatol Online J 2012;18:2.

14 Segagni D, Tibollo V, Dagliati A, et al. The ONCO-I2b2 project: integrating biobank information and clinical data to support translational research in oncology. Stud Health Technol Inform 2011;169:887–91.

15 Anderson N, Abend A, Mandel A, et al. Implementation of a deidentified federated data network for population-based cohort discovery. J Am Med Inform Assoc 2012;19(e1):e60–7.

16 Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009;16:624–30.

17 DHHS. Understanding HIPAA Privacy: Special topics: Research 2002 [updated June 5, 2013]. http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/research/

18 Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity 1997. http://www.whitehouse.gov/omb/fedreg_1997standards/

19 Kalenderian E, Ramoni RL, White JM, et al. The development of a dental diagnostic terminology. J Dent Educ 2011;75:68–76.

20 Schwartz S, Stege D. Tooth numbering systems: a final choice. Ann Dent 1977;36:99–106.

21 Schleyer TK, Ruttenberg A, Duncan W, et al. An ontology-based method for secondary use of electronic dental record data. AMIA Jt Summits Transl Sci Proc 2013;2013:234–8.

22 Kalenderian E, Walji M, Ramoni RB. "Meaningful use" of EHR in dental school clinics: how to benefit from the U.S. HITECH Act's financial and quality improvement incentives. J Dent Educ 2013;77:401–15.

23 Centers for Medicare & Medicaid Services. Medicare Dental Coverage 2013 [updated 19 November 2013]. http://cms.hhs.gov/Medicare/Coverage/MedicareDentalCoverage/index.html

24 Bebinger M. Put Back The Teeth? Why We Separate Dental And Medical Care. WBUR's Common Health Reform and Reality [Internet]. 04/12/2014. http://commonhealth.wbur.org/2014/02/dental-medical-health-care

25 Powell V, Din F. Rationale and need to articulate medical and dental data. In: Powell V, Din F, Acharya A, eds. Integration of medical and dental care and patient data. 3. London: Springer, 2012:1–24.

26 International HLS. HL7/ASTM Implementation Guide for CDA® R2 -Continuity of Care Document (CCD) Release 1: Health Level Seven International; http://www.hl7.org/implement/standards/product_brief.cfm?product_id=6 [cited 04/12/2014].

27 Ribick A, Stanford S. Health Level Seven and the American Dental Association Sign Agreement to Develop Joint Healthcare IT Standard Initiatives 2009. http://www.hl7.org/documentcenter/public_temp_75155CE5–1C23-BA17-0CC4041C0AA87625/pressreleases/hl7_press_20090902.pdf