UC Davis

UC Davis Previously Published Works

Title

Toward Data-Driven Radiology Education-Early Experience Building Multi-Institutional Academic Trainee Interpretation Log Database (MATILDA).

Permalink

https://escholarship.org/uc/item/3t7319w6

Journal

Journal of digital imaging, 29(6)

ISSN

0897-1889

Authors

Chen, Po-Hao
Loehfelm, Thomas W
Kamer, Aaron P
et al.

Publication Date

2016-12-01

DOI

10.1007/s10278-016-9872-2

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

J Digit Imaging (2016) 29:638–644
DOI 10.1007/s10278-016-9872-2

Toward Data-Driven Radiology Education—Early Experience Building Multi-Institutional Academic Trainee Interpretation Log Database (MATILDA)

Po-Hao Chen[1] · **Thomas W. Loehfelm**[2,3] · **Aaron P. Kamer**[4] · **Andrew B. Lemmon**[2] · **Tessa S. Cook**[1] · **Marc D. Kohli**[3]

Published online: 4 March 2016
© Society for Imaging Informatics in Medicine 2016

Abstract The residency review committee of the Accreditation Council of Graduate Medical Education (ACGME) collects data on resident exam volume and sets minimum requirements. However, this data is not made readily available, and the ACGME does not share their tools or methodology. It is therefore difficult to assess the integrity of the data and determine if it truly reflects relevant aspects of the resident experience. This manuscript describes our experience creating a multi-institutional case log, incorporating data from three American diagnostic radiology residency programs. Each of the three sites independently established automated query pipelines from the various radiology information systems in their respective hospital groups, thereby creating a resident-specific database. Then, the three institutional resident case log databases were aggregated into a single centralized database schema. Three hundred thirty residents and 2,905,923 radiologic examinations over a 4-year span were catalogued using 11 ACGME categories. Our experience highlights big data challenges including internal data heterogeneity and external data discrepancies faced by informatics researchers.

✉ Po-Hao Chen
chenp@uphs.upenn.edu

[1] Department of Radiology, Hospital of the University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104, USA

[2] Department of Radiology, Emory University Hospital, Atlanta, GA, USA

[3] Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA

[4] Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA

Keywords Radiology training · Residency · Case log · Education · Database · Big data · Analytics · ACGME

Background

Radiology education, like graduate medical education at large, is conducted primarily using an apprenticeship model. By independently interpreting an imaging study before reviewing with staff radiologists, residents gain experience and knowledge unique to that garnered from reading textbooks [1]. Therefore, the number of interpreted cases is an important surrogate for experience and competence during training.

Experience-based learning in clinical radiology training is important in national requirements as well as hospital credentialing processes. The Accreditation Council of Graduate Medical Education (ACGME) requires that radiology residency programs maintain case logs for 11 categories of examinations as markers for clinical exposure and sets minimum requirements in each category [2]. For example, the ACGME requires that a graduating fourth year radiology resident interpret at least 1900 chest radiographs during training. The Mammography Quality Standards Act requires 240 mammographic interpretations during a 6-month interval within the last 2 years of residency [3]. Other professional organizations such as the American Heart Association/American College of Cardiology Foundation also set minimum interpretation requirements for standardized levels of training [4]. Additionally, some hospitals require that job applicants submit case logs documenting the studies they have interpreted or the procedures they have performed as proof of experience.

Furthermore, radiology training has undergone dramatic changes. The new core examination in 2013 brought forth the necessity for new clinical curricula and preparatory materials [5, 6]. Additionally, the pervasiveness of mobile technology

has changed radiology education [7]. Commercial and homegrown analytic solutions are growing within radiology. However, these tools tend to be focused on clinical productivity and relative value unit (RVU) generation. Limited solutions exist for analytics focused on radiology training. In addition, such efforts are primarily limited to a single institution, precluding the possibility of comparing curricular differences between sites. We describe a multi-institutional academic trainee interpretation log database (MATILDA), created by consolidating data from three US diagnostic radiology residency programs.

Methods

We created MATILDA to study radiologic examinations interpreted by trainees at three American radiology residency programs, one in the South, one in the Midwest, and a third in the Northeast. Each examination was stored with a unique identifier combining the institution's name with a hashed institutional case identifier using the MD5 algorithm. Each entry contained a de-identified but unique resident code, graduation class, case timestamp, and case descriptor (Table 1). These protocols were reviewed by the institutional review board at each of the three institutions, and a waiver was issued.

Each examination was assigned to one of the 11 categories tracked by ACGME or to a 12th group labeled "Other." If available, Current Procedural Terminology (CPT) codes were used to ensure proper assignment to the ACGME category. Studies where CPT codes were unavailable were categorized manually. We attempted to map to RadLex PlayBook procedure codes but opted to implement it as an optional field after realizing it was outside the scope of our project focusing on ACGME categories.

We included examinations and procedures dictated by radiology residents between 1 July 2010 and 30 June 2014. Each of the three institutions used various mechanisms to retrieve interpretation data, including radiology information system (RIS) queries and electronic medical record system data warehouses. Data from each institution were adapted to conform to the MATILDA schema. Figure 1 shows a schematic depiction of the integration process.

Site no. 1 At site no. 1, a different radiology information system was in place for each of several member hospitals. A customized case log system was built to centralize case logs for all the residents. The centralized case log was initially maintained using Access (Microsoft Corporation, Seattle, WA) and subsequently converted to an open-source server-based program MySQL (Oracle Corporation, Santa Clara, CA). Site 1 periodically updated its database by querying the individual member hospitals' medical record databases and then consolidating the results. A unique identifier in the central database associated each radiology resident with their separate, member hospital-specific identifiers, and data were merged in the centralized case log. An Access query was performed to transform data to the MATILDA specification. Data from the Veterans Affairs Medical Center was inaccessible; this excluded less than 1 % of the total number of studies dictated during residency.

Site no. 2 Site no. 2 had multiple radiology information system (RIS) instances from three different vendors (Siemens, Cerner, Fuji), each with unique message specifications. Health-Level Seven (HL7) feeds from the three RIS instances were integrated into a unified data warehouse as part of a previous project. Trainee identity was mapped at the time of population in the data warehouse. A Ruby script mined the data warehouse using a lookup table for each of the radiology trainees and assigned ACGME categories to interpreted exams based on stored CPT and/or local procedure codes. This data was then exported as a MATILDA-compatible comma-separated values (CSV) file. Due to data access restrictions, site no. 2 was unable to incorporate resident dictations from rotations at the local Veterans Affairs Medical Center into the case logs. Further, some of site no. 2's outreach MRI examinations were also excluded.

Site no. 3 Site no. 3 utilized an open-source application for resident-centric case logging and analysis. The radiology information system (RIS; General Electric Centricity RIS-IC, Barrington, IL) was queried based on resident identification numbers and shifts, and the results separately maintained in a secured MySQL (Oracle Corporation, Santa Clara, CA) database optimized for residents. The software used an incremental update algorithm that updated the database hourly. A unique code for the entirety of their 4-year training period de-identifies the residents while maintaining the ability to longitudinally correlate performance. A query of this MySQL database was performed to export a database satisfying the MATILDA specification. Data from the site's affiliated children's hospital and Veterans Affairs Medical Center were unavailable for incorporation, estimated to affect approximately 1 % of the total volume of interpreted cases.

Consolidation

Finally, we combined the compatible case logs from each of the three institutions into the aggregated database, maintained

Table 1 Multi-institutional academic trainee interpretation log database schema

Column name	Data type	Length	Description	Example
txtAccNumHash	Text	35	Institution followed by 128-bit hashed identifier (e.g., accession number)	SITE-e8919413fd481a5a33448ebe47aa46e2
txtResidentID	Text	7	Institutional-specific resident ID	SITE153
txtStudyName	Text	255	Study description	CT abdomen and pelvis with IV contrast
numGradYear	Integer	4	Four-digit year	2013
dteTimestamp	Date time	N/A	In UTC format	2013-02-26T13:35-05:00
numCPT	Text	255	Best-fit CPT codes, if available	74177
txtRPID	Text	8	Best-fit RadLex Playbook ID, if available	RPID860
txtModality	Text	4	DICOM-compatible modality descriptor	CT
txtACGMECat	Integer	2	Database internal designation for ACGME categories for minimum procedural numbers	2

in Microsoft Access. Data analysis was performed in Access, Microsoft Excel, and R version 3.2.0 [8].

Results

Three hundred thirty residents in three academic training programs interpreted 2,905,923 radiology examinations over the 4-year span that was included in our database. The breakdown of resident interpretations by training program and by ACGME categorization is detailed in Table 2. Although all radiology residents are subject to the same ACGME case log requirements, training programs differ widely in volume and mix of radiologic examinations, creating variable educational experiences for radiology residents across the country.

Furthermore, there is marked heterogeneity in vendors as well as hospital-specific customizations for radiology information systems across the nation. Informatics personnel spent approximately a total of 40 h across the three sites in creating the unified MATILDA dataset, with site no. 1 spending 15 h; site no. 2, 10 h; and site no. 3, 15 h. The bulk of the time spent was in normalizing trainee identifiers and matching local procedure codes to ACGME categories.

Discussion

Creating a multi-institutional academic trainee interpretation log database requires balancing between compatibility and complexity. Differences in examination protocols and database storage formats make cross-institutional comparison of residency education experience in radiology difficult. A centralized schema is necessary to ensure cross-institutional comparability, and MATILDA allowed trainee interpretation logs across three institutions to be consolidated and compared.

Maintaining such a database requires a flexible schema that allows each institution to input data while providing sufficient granularity and volume of data for analysis, while still preserving resident anonymity. Although MATILDA is designed to analyze resident case logs, designing, building, and maintaining the multi-institutional database requires solutions to many challenges the practice of radiology must tackle in the age of "big data."

Although the concept of big data is nebulous, a systematic healthcare literature review by Baro et al. reports several components as particularly relevant to healthcare: volume, variety, velocity, and veracity [9]. Additional characteristics have also been proposed: variability, visualization, and value [10–12]. These terms are sometimes collectively referred to as the "V's of big data." Examples of healthcare related big data include electronic medical records, images, and diagnostic reports from radiology and pathology, national utilization databases, social media, and biologic "–omics" [10, 13, 14].

Despite the emergence of data science in healthcare and in radiology, there is a paucity of literature exploring multi-institutional data analytics in radiology education. By assembling case log information from three institutions across 4 years and studying a total of 330 residents, MATILDA is uniquely poised to answer questions about residency education. A brief review of MATILDA shows that an average radiology resident interprets approximately 16,800 examinations throughout residency, while the minimum required by the ACGME is 3500 [2]. It should be noted that the ACGME specifications are not necessarily designed to capture the complete resident experience, and instead measure a few types of studies that might be considered as surrogates for the overall experience. The ACGME might use such data to determine whether a program has sufficient case volume to support additional training spots, for example, and not necessarily to certify that any individual resident has received adequate training. However, the extent to which the ACGME minima underestimate the total resident experience was not clear until our study. Our data indicate that approximately 50 % of resident work is not captured by the 11 ACGME categories, and

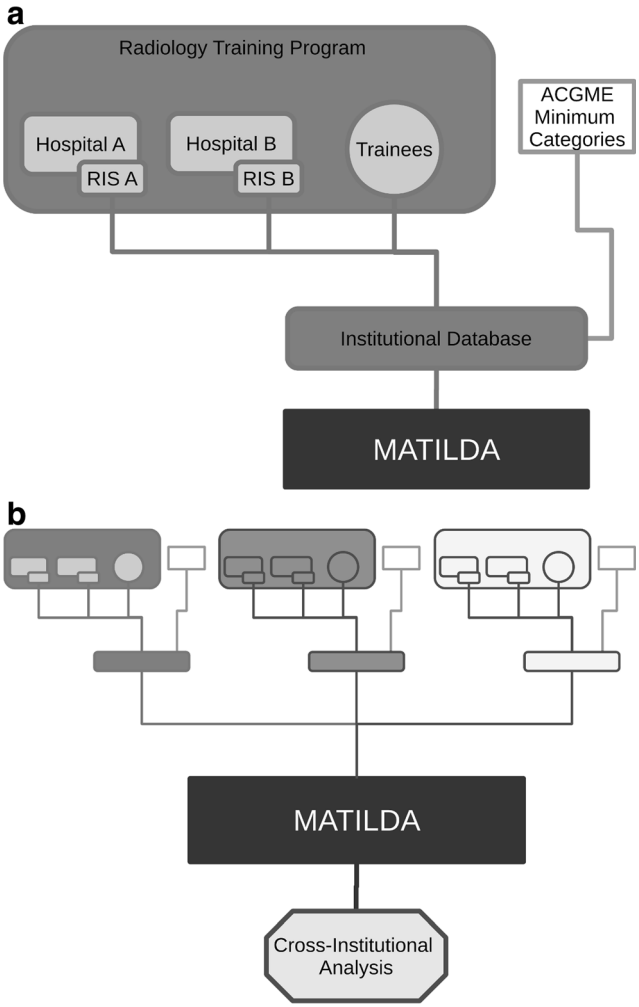


Fig. 1 Schematic depiction of the consolidated database creation. **a** Each radiology training program combines data as appropriate, and **b** each of the institutional databases is consolidated using the common database schema, enabling quantitative cross-institutional analysis

that the resident experience is not evenly distributed among or within training programs. An in-depth analysis of other residency training trends is forthcoming in separate publications.

The database creation process required management of both inter- and intra-institutional data heterogeneity. Our experience agrees with existing literature suggesting that combining data from disparate systems is a primary "big data" problem in healthcare [12]. Unfortunately, a paucity of data in the literature exists to quantify the degree of heterogeneity in a practical fashion. Using the amount of time spent as a surrogate marker for complexity, our team spent a total of 40 h across three sites to reconcile inter-institutional differences based on MATILDA specification to merge our data.

However, in order to create each institutional database, we parsed sources from the larger institutional RIS at each site for relevant records related to resident interpretations. Although each of the three sites maintains only one residency program, each training site consists of multiple member hospitals, many with separate electronic medical records. Therefore, each of the three sites managed intra-institutional data heterogeneity in a different fashion. Site no. 1 relied on periodic manual querying for resident data in its various, otherwise incompatible, databases and then assembled the results into a centralized database. While some of these steps can and were subsequently automated, there is still a minimum amount of manual work required to update and maintain the system. Individual sites use different medical record vendors, so site no. 2 uses a database warehouse which provides centralized access to the information. Site no. 3 uses a single vendor for the RIS database and uses a backup RIS server that contains comprehensive interpretation records. Then, the veracity of each database is confirmed at the institutional level—by each of the individual sites, which involves both manual and automated means. We estimate that we expended approximately 60–100 person-hours at each site to perform these tasks.

Solving a radiology practice's "big data" variety problem turned out to have long-reaching benefits, enabling the creation of other technological solutions to improve residency education. For instance, site no. 1 combined case log data with resident scheduling software (QGenda, Incorporated, Atlanta, GA) to correlate between a resident's progression through residency and satisfaction of ACGME requirements. Site no. 2 implemented an open-source solution connecting to its data warehouse to allow its residents timely access to volume-based case log data. Site no. 3 implemented the same open-source solution, but by querying a backup RIS server rather than sourcing information through a data warehouse.

Furthermore, building a centralized database case log also enabled each institution to implement case log visualization for residents. The value of these tools arises from simultaneous visualization of numerous variables greater than the

Table 2 Volume ACGME categories for three academic institutions from 1 July 2010 to 30 June 2014

ACGME categories	Site no. 1	Site no. 2	Site no. 3	Total
0. No category applies	637,031	498,471	321,130	1,456,632
1. CXR	285,743	276,716	160,883	723,342
2. CT AP	67,591	74,010	46,688	188,289
3. CTA/MRA	35,202	9755	15,994	60,951
4. Biopsy/drainage	3684	13,309	3623	20,616
5. Mammography	40,214	39,180	32,522	111,916
6. MRI body	26,079	6758	17,912	50,749
7. MRI knee	7805	5838	2876	16,519
8. MRI spine	17,933	12,969	11,436	42,338
9. PET	12,092	7549	12,448	32,089
10. US AP	72,385	51,283	16,954	140,622
11. MRI brain	27,769	21,723	12,368	61,860
Total	1,233,528	1,017,561	654,834	2,905,923

capacity of human cognition [12]. Traditionally, residents have two options for case logging. They may manually record cases, or they may wait for semi-annual summary reports from their training programs. Manual recording is tedious and prone to inaccuracy. The literature on ACGME surgical resident case logs suggests underreporting when using manual submission, with only half of the residents surveyed recording major procedures and only 13 % logging minor procedures [15, 16]. On the other hand, waiting for the semi-annual report may preclude proper intervention should a trainee fall behind case log minimum numbers. The efforts invested to create a centralized database at each of the three sites allowed all three institutions to implement visualization tools for on-demand access of personalized case logs. In two of the sites, the same open-source solution was implemented by accessing each institutional residency database [17]. Sample screenshots of the front-end visualization interface at different sites are shown in Fig. 2.

To our knowledge, MATILDA is the first multi-institutional, educationally focused radiology case logging database. Although it shares many similarities with single-institution solutions, integrating externally heterogeneous datasets creates additional challenges. The tasks of unifying data format, maintaining unique identifiers, and mapping site-specific data to a unified scheme are known problems in big data analytics [13, 18]. In our case, one challenge involves resolving timestamps as the site cohort spans multiple time zones. To eliminate potential heterogeneity in examination timestamps, we used coordinated universal time (UTC) for all dates. A second challenge requires elimination of externally meaningless data. A hospital's accession number scheme for radiology examination is incompatible for a different hospital system and risks potential duplication in the multi-institutional database. Our schema utilizes a separate unique identifier by applying a hash algorithm to the site name and corresponding accession number to avoid this potential pitfall. Additionally, each site's examination code is unique to each site, often unique to a specific hospital. This problem is solved by requiring each site to assign each examination an ACGME categorization code prior to data unification. The current schema includes optional fields for both CPT and RadLex codes, but we determined that ACGME categories were better for the purposes of studying trainee volumes, as they provided reasonable, higher-level groupings for analysis. For example, while Medicare may want to know whether a chest X-ray included only the standard two views or additional apical lordotic and oblique views, which are captured by separate CPT codes, for our purposes, it is more useful to know the total number of chest X-rays that were interpreted.

Some cross-institutional differences are intrinsic to the heterogeneity of ACGME compliant training. For example, some sites may include both research and clinical residency pathways, with research pathway residents interpreting significantly fewer examinations in their final year of residency. Furthermore, each site approached the new radiology core examination differently, some by incorporating protected off-service time into the curriculum while others did not. To account for some of these differences and allow for post hoc analyses in the future, we included graduation year and site name in the schema.

Nevertheless, a multi-institutional database requires management of several risks. For patient protection, the database contains no identifiable information complying with Health Insurance Portability and Accountability Act (HIPAA). We assigned each interpreting resident a random number to obfuscate his/her identity. Data access was approved by institutional review prior to exchanging data.

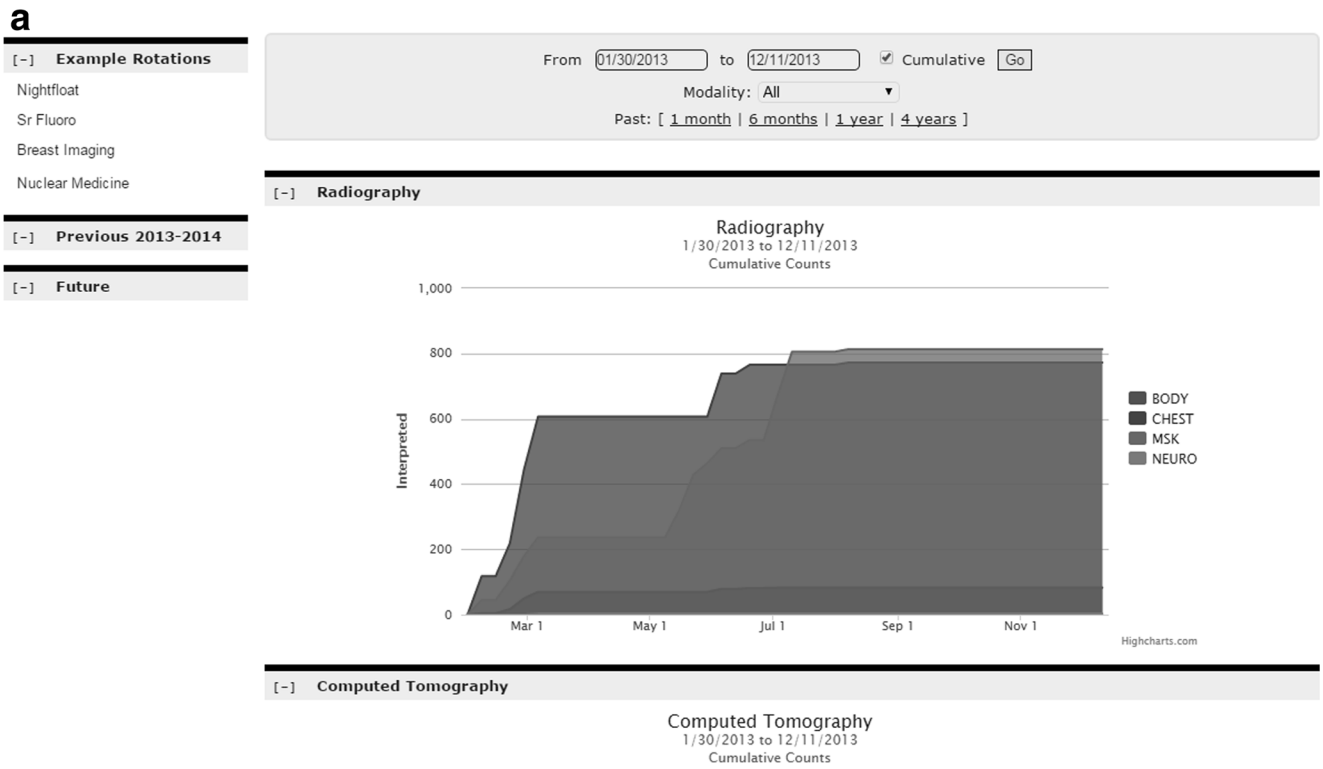
Despite the best efforts, our multi-institutional database has several limitations. First, there is a limitation due to the data quality in the original systems that cannot account for cases when the radiologist who dictated that the procedure report did not designate all trainees as contributors through the dictation software. This sometimes occurs on procedural services such as interventional radiology or with biopsy cases, when multiple trainees contribute to a patient's care but only one dictates the report. Although the degree by which this phenomenon occurs can be quantified by comparing manual procedure logs against MATILDA, we do not have this comparison data. Additionally, not all hospitals in each academic institution provide access to complete electronic records. To the extent that a site is unable to query the radiology information system at a member hospital, resident case volume may be underestimated. For the three participating institutions, we estimate that approximately 1 % of interpretation volume is inaccessible.

Future directions include answering a number of important educational questions using a quantitative approach. For instance, the database allows quantitative measurement of how and when residents meet the ACGME criteria throughout their residency training and potential factors that affect the trajectory. The database can also evaluate whether decreased preparation time for the core examination allowed fourth year residents to return to clinical duty, one of the motivations behind the new core examination [6].

We hope to collaborate with other radiology training programs to augment the quantitative analytic power of the database to shed light on radiology training. Finally, another future direction involves streamlining database integration, a process that still requires some level of manual processing to ensure data integrity at this time.

Conclusion

Our experience building MATILDA required collaborative solutions to several "big data" challenges. A multi-



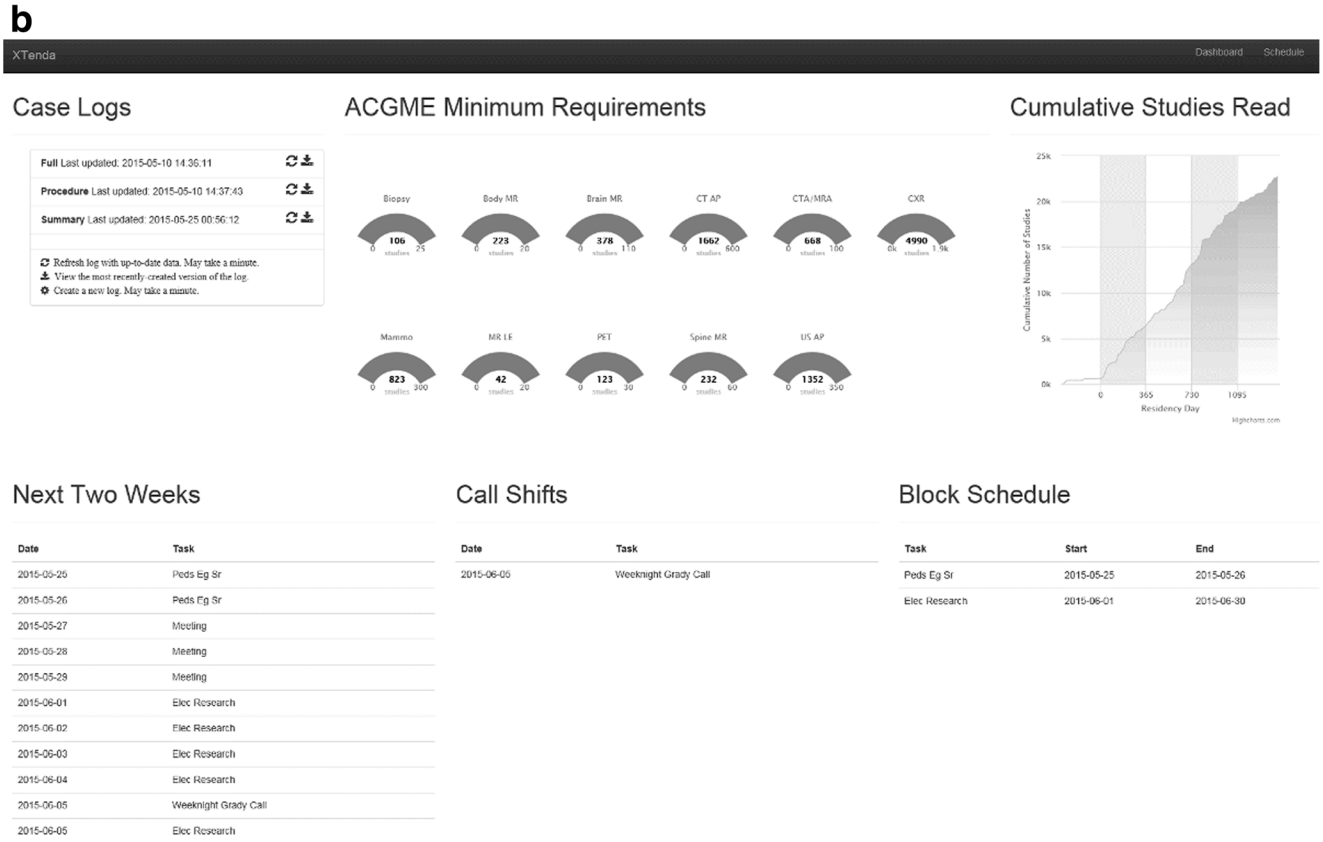


Fig. 2 Sample screenshots depicting resident case log visualization tools from two of the sites contributing to the multi-institutional database. Two of the institutions implemented an open-source solution (**a**), while the third institution created a different interface (**b**)

institutional case log database requires delicate balance between maintaining the compatibility among high variability of radiology information systems and the complexity necessary for meaningful research. Using a standardized case log schema, the database enables quantitative analysis of diagnostic radiology residency training.

References

1. Williamson KB, Gunderman RB, Cohen MD, Frank MS: Learning Theory in Radiology Education1. Radiology 233(1):15–18, 2004
2. Accreditation Council for Graduate Medical Education. Diagnostic Radiology Case Log Minimums [Internet]. [cited 2014 Jul 1]. Available from: https://www.acgme.org/acgmeweb/Portals/0/PFAssets/ProgramResources/420_DR_Case_Log_Minimums.pdf
3. Compliance Guidance: The Mammography Quality Standards Act Final Regulations: Preparing For MQSA Inspections [Internet]. U.S. Food and Drug Administration. 2001 [cited 2014 Jul 4]. Available from: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094420.htm
4. Budoff MJ, Cohen MC, Garcia MJ, Hodgson JM, Hundley WG, Lima JAC, et al: ACCF/AHA clinical competence statement on cardiac imaging with computed tomography and magnetic resonance. Circulation 112(4):598–617, 2005
5. Nachiappan AC, Wynne DM, Katz DP, Willis MH, Bushong SC: A proposed medical physics curriculum: preparing for the 2013 ABR examination. J Am Coll Radiol JACR 8(1):53–57, 2011
6. DeStigter KK, Mainiero MB, Janower ML, Resnik CS: Resident clinical duties while preparing for the ABR core examination: position statement of the Association of Program Directors in Radiology. J Am Coll Radiol JACR 9(11):832–834, 2012
7. Bhargava P, Lackey AE, Dhand S, Moshiri M, Jambhekar K, Pandey T: Radiology education 2.0–on the cusp of change: part 1. Tablet computers, online curriculums, remote meeting tools and audience response systems. Acad Radiol 20(3):364–372, 2013
8. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing
9. Baro E, Degoul S, Beuscart R, Chazard E: Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Res Int 2015:639021, 2015
10. Sukumar SR, Natarajan R, Ferrell RK: Quality of Big Data in health care. Int J Health Care Qual Assur 28(6):621–634, 2015
11. Vaitsis C, Nilsson G, Zary N: Big data in medical informatics: improving education through visual analytics. Stud Health Technol Inform 205:1163–1167, 2014
12. Caban JJ, Gotz D: Visual analytics in healthcare—opportunities and research challenges. J Am Med Inform Assoc JAMIA 22(2):260–262, 2015
13. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics [Internet]. 2015 Dec [cited 2015 Jul 26];8(1). Available from: http://www.biomedcentral.com/1755-8794/8/33
14. Fernández-Luque L, Bau T: Health and social media: perfect storm of information. Healthc Inform Res 21(2):67–73, 2015
15. Gow KW, Drake FT, Aarabi S, Waldhausen JH: The ACGME case log: general surgery resident experience in pediatric surgery. J Pediatr Surg 48(8):1643–1649, 2013
16. Salazar D, Schiff A, Mitchell E, Hopkinson W: Variability in Accreditation Council for Graduate Medical Education Resident Case Log System practices among orthopaedic surgery residents. J Bone Joint Surg Am 96(3), e22, 2014
17. Chen PH, Chen YJ, Cook TS: Capricorn—A Web-Based Automatic Case Log and Volume Analytics for Diagnostic Radiology Residents. Acad Radiol 22(10):1242–1251, 2015
18. Hendler J: Data Integration for Heterogenous Datasets. Big Data 2(4):205–215, 2014