J Digit Imaging (2016) 29:331–336
DOI 10.1007/s10278-015-9817-1

Datafish Multiphase Data Mining Technique to Match Multiple Mutually Inclusive Independent Variables in Large PACS Databases

Brendan P. Kelley[1] · **Chad Klochko**[1] · **Safwan Halabi**[1] · **Daniel Siegal**[1]

Published online: 16 November 2015

Abstract Retrospective data mining has tremendous potential in research but is time and labor intensive. Current data mining software contains many advanced search features but is limited in its ability to identify patients who meet multiple complex independent search criteria. Simple keyword and Boolean search techniques are ineffective when more complex searches are required, or when a search for multiple mutually inclusive variables becomes important. This is particularly true when trying to identify patients with a set of specific radiologic findings or proximity in time across multiple different imaging modalities. Another challenge that arises in retrospective data mining is that much variation still exists in how image findings are described in radiology reports. We present an algorithmic approach to solve this problem and describe a specific use case scenario in which we applied our technique to a real-world data set in order to identify patients who matched several independent variables in our institution's picture archiving and communication systems (PACS) database.

Keywords Data mining · Databases · Image database · Imaging informatics · PACS · Software design · User interface

Background

The Datafish Project was initially created out of necessity to determine if a sufficient number of patients existed within our institution's PACS database to retrospectively study the potential prognostic significance of bone marrow edema-like signal abnormalities seen on MRI knee following acute trauma. Institutional review board approval was obtained, and informed consent was waived for this retrospective study. We needed to quickly and accurately ensure that there were enough patients left with bone marrow edema-like signal abnormalities in their knee on MRI after trauma to reach statistical significance after excluding all the patients with any of the various confounding variables that could have contributed to the clinical and radiologic outcomes being assessed in the study design. This meant that we needed to exclude patients with acute fractures or advanced osteoarthritis on their initial radiograph, as well as exclude patients with ACL tears on their initial MRI knee. This also required that we exclude patients with "clinically significant" meniscal pathology, which required a focused review of the electronic medical record when meniscal signal changes or discrete tears were reported on MRI knee. Given these multiple exclusion criteria, it was crucial that we started with the largest possible cohort of patients in the PACS database with the necessary combination of knee radiographs and MR exams that occurred at the appropriate proximity in time relative to their acute knee trauma.

Fortunately, our institution recently reached the informatics milestone of generating a database with 10 years' worth of searchable radiology data (currently over 9 million records). Unfortunately, the existing data mining software at our institution did not have advanced search features capable of identifying the subset of patients with the right combination of imaging studies, which consisted of patients with an initial knee radiograph, and an MRI knee within 1 month, and at least one follow-up knee radiograph in the next 1–2 years. This particular set of imaging studies and proximity in time was a fundamental inclusion criterion for the study given that the central research objective was to correlate imaging

✉ Brendan P. Kelley
brendank@rad.hfh.edu

[1] Department of Radiology, Henry Ford Hospital, 2799 W Grand Blvd, Detroit, MI 48202, USA

findings with clinical outcomes after acute knee trauma. Identifying the specific patient population we wanted to study within the PACS database became a real challenge, and out of necessity, the Datafish Project was created to find a solution. We present the novel algorithmic approach that emerged as the solution to this problem which is designed around our prototype Datafish software and consists of three distinct data mining phases: (1) data acquisition from our institution's pre-existing data mining software (Softek Solutions Inc., Prairie Village, KS), (2) automated filtering of the exported data in Excel using a custom data mining algorithm written with Visual Basic for applications (VBA, Microsoft, Redmond WA), and (3) manual sorting using a VBA-coded customized user interface (UI) to produce a final patient list.

It is worth noting that sophisticated data mining tools capable of accomplishing a task like the one presented here are already well established in other industries and are being extensively used by financial institutions, insurance providers, and social media companies to continuously identify complex patterns and trends in large real-world data sets [1–3]. Data mining in healthcare is noticeably underdeveloped compared to these industries, and effective use of current medical data mining tools often requires a customized end user solution like Datafish to bridge the functional gap to aggregating, organizing, and analyzing specific types of medical information. Impressive progress has already been made in certain areas of medicine related to cancer genomics [4], molecular biology [5], pharmacology [6], and the epidemiology of chronic disease [1]. Most of this early success has been achieved through open collaboration within standardized online data sharing communities [4–8] that modify existing mathematical models with complex analytical algorithms and rigorous statistical filtering techniques to establish a computational data mining method which often combines predicted and experimentally determined results using a Bayesian framework [3, 5, 6, 8]. Genome-wide association studies are a particularly promising development in medical data mining because they can analyze billions of data points, and in radiology, this big data methodology has already proven proficient at using MRI databases for brain-wide association studies (whole-brain voxel pairwise analysis) capable of identifying the key functional and anatomical differences that form the neural bases for cognitive disorders such as autism, obsessive compulsive disorder, ADHD, and schizophrenia [8]. We were unable to find open source or commercially available data mining tools to meet the specific needs of our project, and in the tradition of open collaboration, we present our initial experience with the custom solution developed in the Datafish Project for those wishing to identify patients with multiple mutually inclusive variables across multiple different medical databases.

Methods

The general concepts behind the Datafish Project's data mining technique are outlined in Figs. 1, 2, and 3 which correspond to the three unique phases of our method. Phase 1 in the use case scenario consisted of generating two large "modality-specific" patient lists from our PACS database using only minimal search criteria. These two separate lists consisted of all patients at our institution who had received an MRI or radiograph of the knee over a 10-year period. This search revealed 26,000+ MRI exams and 400,000+ radiograph results which met this basic inclusion criterion. These data sets were then extracted from the database and imported into Excel under separate sheets. In phase 2, the individual data sets were cross-referenced according to more complex filtering criteria using a custom data mining algorithm coded in VBA. In order to be included in phase 3, a patient needed to be on both lists (MRI, radiograph), with the additional requirement that they had an MR and radiograph within 1 month of each other and at least one follow-up radiograph within 1–2 years. This automated filter coded in VBA produced a single patient list of approximately 700 patients with the necessary imaging studies. Manually sorting the patient lists by hand according to this complex inclusion criterion would have required a lot of time and manpower, specifically because the added complexity of filtering by multiple mutually inclusive independent variables prohibited a simple Boolean search of the database in phase 1.

With the patient list now substantially smaller and more specialized after automated filtering in phase 2, the manual

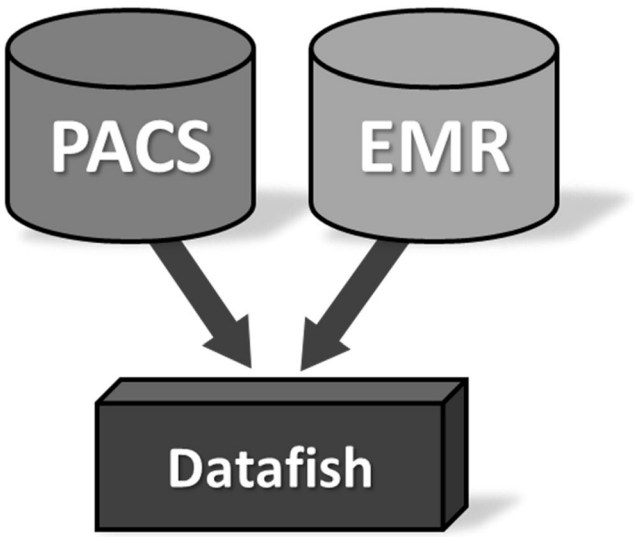

Fig. 1 Phase 1—data acquisition. Obtain individual data sets from any of the various medical databases and import them into Datafish. Chances are these databases do not directly communicate and are not centralized for single access data acquisition. Phase 1 allows the user to "semi-centralize" data from multiple databases across the health system in order to perform complex searches using all available information. In the use case scenario, only PACS data sets were obtained

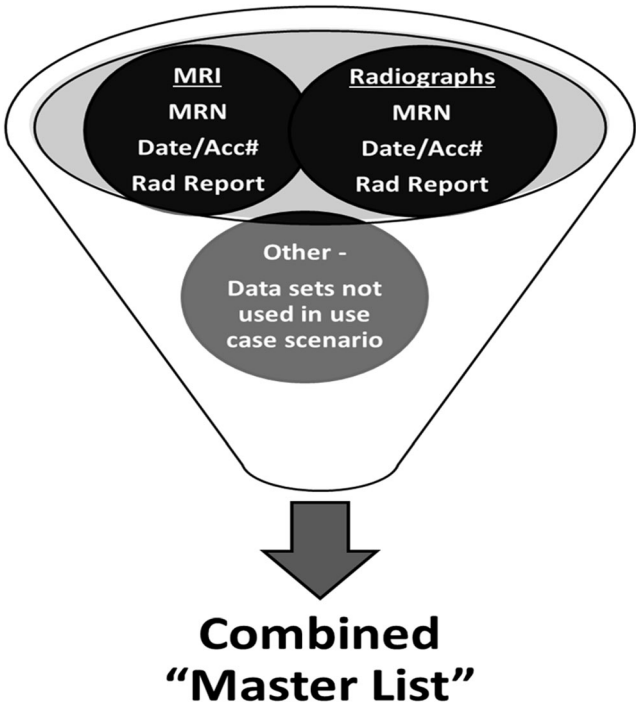

Fig. 2 Phase 2—custom data mining algorithm. Apply automated filter algorithm within Datafish to identify patients that satisfy overlap criteria across the individual data sets. The custom algorithm uses a brute force iterative matching technique and can be tailored to identify patients according to any number of mutually inclusive independent variables. Phase 2 allows the user to substantially reduce the data set through computational analysis and produces a single patient list that is highly specialized

sorting phase was initiated in phase 3 with a user-dependent filtering of the remaining 700 patients according to more subtle inclusion criteria that could not be written into a VBA-encoded data mining algorithm. The more nuanced filtering criteria applied in this use case scenario related to the radiology reporting of knee pathology and required a technical knowledge of how imaging features within MR and radiograph exams were described by radiologists, with an understanding that some variation still exists in the current age of non-standardized radiology reporting. This manual sorting phase was expedited by the user-form capabilities within VBA, which allows for generation of a custom user interface to display data as desired. A form was created that allowed for quick selection of the presorted radiology reports (see Fig. 3 for an example of the user interface applied in the use case scenario which is displayed with templates modified from the RSNA Radiology Reporting Initiative [9]). The MR report and plain film reports were displayed side by side in the user interface, which decreased the time required for manual analysis by the end user. Buttons were added to the user interface to allow for a quick inclusion or exclusion of patients based on their radiology reports. A separate button was included to flag cases that needed additional chart review. The end result of phase 3 was displayed in an Excel spreadsheet which contained the finalized patient list with the corresponding radiology reports for their knee radiographs and MR exams.

Results

The initial application of this data mining technique proved to be very effective and resulted in a data set of approximately 50 patients from a list of approximately 26,000 MRIs and 400,000 radiographs. Doing this search manually would not have been feasible for a variety of reasons, especially given the inherent challenges of searching a PACS database which contains broad variation in the language used to report imaging findings. Our multiphase technique minimizes the current limitations that exist in the age of non-standardized radiology reporting by separating the data mining process into three unique phases. Another key feature of our Datafish technique is that it minimizes the need to search the entire database for keywords that are often pertinent positive/negatives for a given study and are therefore mentioned in some capacity in the majority of the associated radiology reports. In the use case scenario, phase 1 was quick and easy because the central database could be searched for modality-based data sets (MRI, X-ray) that could be exported within minutes. In phase 2, the data mining algorithm encoded within VBA identified all patients that met the custom filter criteria and copied them into a separate Excel sheet to be included in phase 3. One particularly nice feature was that all of the patient's studies were automatically copied and included for manual review in phase 3 within the visual user interface (Fig. 3).

The Datafish VBA program ran the data mining algorithm in phase 2 for approximately 36 h across nearly 500,000+ imaging studies on a 6-core 3.2 GHz processor with 8 GB of DDR 1600 memory and Windows 7 (Microsoft Redmond, MA) and generated an Excel spreadsheet of 700 patients who met the multiple mutually inclusive variables written into the VBA code. The final process of manual sorting in phase 3 was facilitated by the custom user interface and took a single end user 12 h over five sessions to identify approximately 50 patients from the list of 700 patients who were ideal candidates to study the potential prognostic significance of specific bone marrow edema-like signal abnormalities seen on MRI knee following acute trauma. The relevant imaging studies for this final patient list will be uploaded to a research PACS at our institution for a focused image review by three musculoskeletal radiologists who will be blinded to the original reporting and will characterize the bone marrow edema patterns on MRI knee as well as the interval changes seen between the initial and follow-up radiographs. Sufficient overlap among radiologists will be ensured to calculate interobserver variability. Following the image review, the results will be assessed for statistical significance, and the findings will be submitted for publication in a separate manuscript which will include a discussion of any possible bone marrow edema patterns that were shown to correlate with poor clinical outcomes.

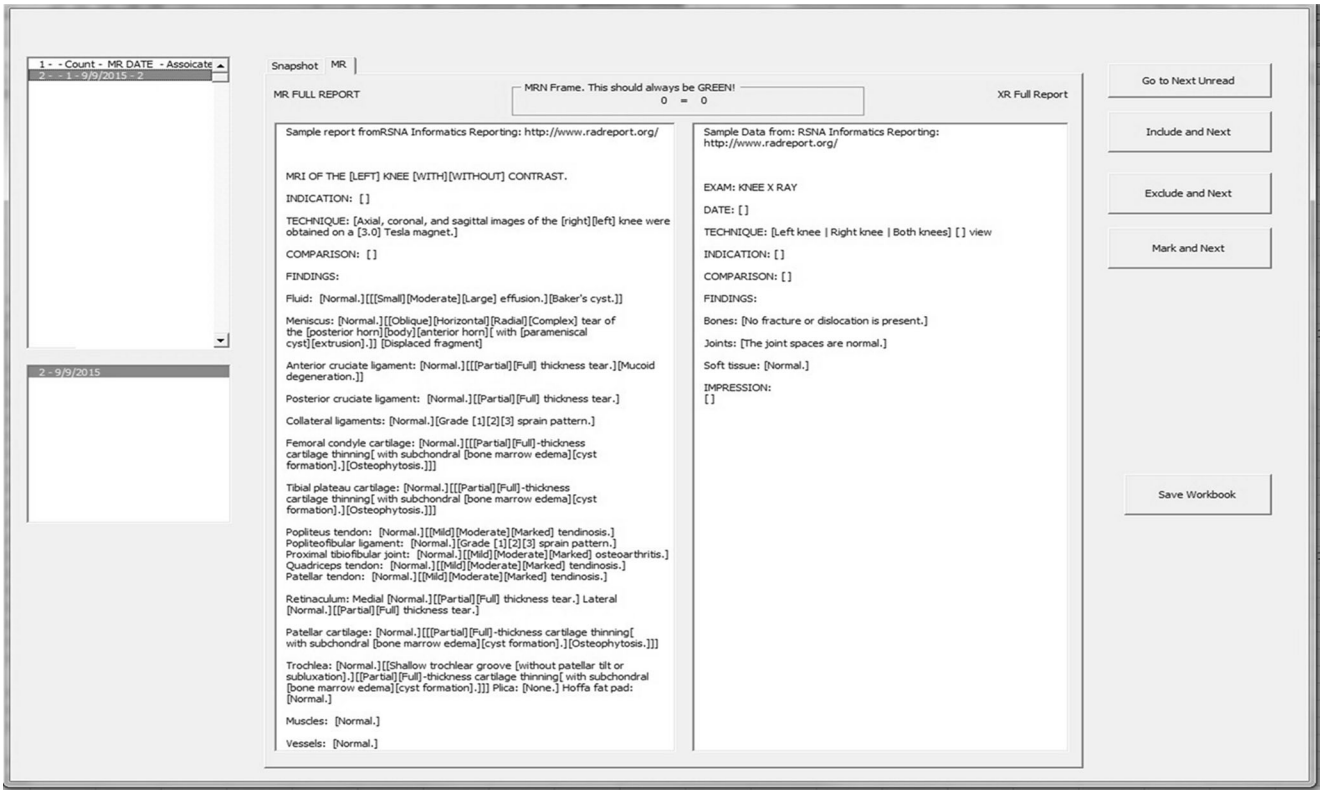

Fig. 3 Phase 3—manual sorting. Display search results in a custom user interface that organizes the specialized data set as desired by the user to facilitate rapid selection of patients according to user-dependent criteria that are often more nuanced and require specialized end user knowledge that cannot be reliably encoded into an automated filter. Phase 3 allows the user to review the search results for accuracy while also producing a final list that represents a highly specific cohort of patients who meet the necessary inclusion/exclusion criteria for medical research

Algorithmic data mining of large PACS databases using this combination of software encoded and end user filtering to identify highly specialized patient populations to study has the potential to significantly expand the complexity and clinical utility of large retrospective research projects at our institution.

Discussion

VBA within the context of Excel allowed very rapid prototyping of data mining algorithms and ultimately produced a program that could faithfully navigate a data set of close to 500,000 patient records. In our initial use case scenario, Excel acted as the database with VBA allowing rapid access, modification, and comparisons of data. This infrastructure was the foundation of the code and contributed to the speed of development within the Datafish Project. This platform within Excel was chosen because our primary software engineer had experience coding in the VBA and because Excel provided basic functionality that would have to have been recreated if another programming solution was utilized. This approach proved useful in the early stage of the Datafish Project, but this approach will not be adequate moving forward. One of the most significant limitations we encountered in VBA was that there is no element of natural language processing in VBA. Also, any data that is automatically sorted must be numerical in nature (dates, medical record numbers, exam identifiers, etc.). Filtering limited to numerical criteria can still reduce the number of records that must be manually sorted by the end user, but expanding automated filtering beyond numerical criteria would significantly ease the burden of manual sorting that plagues large data mining projects. Another limitation of programming within VBA is that the solution must be hard coded for the needs of each particular project. While the concepts and designs discussed are flexible, each VBA-encoded data mining algorithm must be tailored to the unique specifications of each project. There are additional interoperability issues that stem from having to access the program within Excel. Once the program is created, it exists only within the Excel file which contains all of the data. If that file is emailed to three different people for analysis, and those people work on different parts before returning the file at the end of the day, there are three different Excel files with different parts of the research data. Additionally, if the VBA program is updated on one of those three files, there is no way to distribute that change to the other files automatically.

Further limitations to this VBA approach surround the acquisition and formatting of the source data obtained in phase 1 from our institution's central database, which must be extracted by hand and exported from the database in the form of

multiple smaller comma-separated value (CSV) files. This becomes particularly challenging when trying to export source data from multiple different databases. Once the complete data set is rebuilt within Excel, the custom VBA program expects the data to match a specific format organized by columns and rows within sheets. If the order or content of the information provided in the original set is changed, the program will not function until the VBA code is reformatted. In order for data mining software like this to evolve and become useful beyond its initial purpose, it must eventually be developed in a distributed environment. One of the most basic improvements to this prototype would be to transfer the information from a CSV file into a relational database, which would allow for quicker analysis and sorting. This type of architecture would allow for anyone with access to create queries of the data. These changes could be saved for other users, or further modified to refine the patient population or research question. Working in a distributed environment with an Internet-based solution creates a new set of challenges, and a detailed discussion of these challenges is beyond the scope of this manuscript. Having a database with potential HIPAA information connected to the Internet brings up issues of security. The database would have to be password protected and the data locally encrypted to protect against both physical and electronic compromises. We believe these challenges are inevitable, and the solutions these improvements provide outweigh the obstacles to developing this data mining technique within an independent software platform outside of Excel that is primarily cloud based and exists in a multiuser-distributed environment.

Conclusion

One of the central barriers to initiating large retrospective data mining project is the excessive time and labor costs that must be invested up front to analyze source data within a database to see if there is sufficient information to pursue a research project. Manually completing this "exploratory phase" of large retrospective research projects can take months to years, even with a large research group, and there is no guarantee that the results will be favorable. Many researchers have invested significant time and energy in exploring research ideas that ultimately do not pan out, and this experience can dissuade researchers from pursing similar "big data" project in the future if the initial exploratory process is too time consuming or labor intensive. There is no doubt that big data in medicine is here to stay, and as the size and complexity of medical data continues to expand, it becomes crucial that researchers have the necessary tools to conduct efficient and accurate research despite the enormous amount of information that must be analyzed. Advances in medical informatics and computational analytics will obviously play an important role in how research is conducted moving forward, but it would be foolish to minimize the role of trained physicians and scientists in the research process. It is our belief that a balanced collaboration between computational methods and end user analysis is essential to making meaningful progress in big data research; therefore, we believe it is crucial that software be designed to harness the incredible power of automated filtering techniques while also providing a user friendly interface for physicians and scientists to confirm the accuracy of computer-generated results while also applying their own specialized knowledge to the investigation of complex medical data sets.

The novel data mining technique developed in the Datafish Project rapidly and accurately filters large data sets using a combination of automated and manual filtering methods. The iterative matching techniques employed in the automated phase allow for rapid cross-referencing of multiple independent variables, which would otherwise be impossible with simple Boolean and keyword search alone. Many of the challenges in retrospective data mining will likely be minimized as we move toward standardized radiology reporting, but in the meantime, our prototype software has demonstrated that it can successfully export large data sets and rapidly analyze them through the use of multivariate filters and cross-referencing techniques to trim the list prior to manual verification, which significantly reduces the manual workload. Our initial experience with this method has been promising, and future iterations of our approach will likely benefit from the software being written as a stand-alone program, independent from Excel, with the ability to incorporate data sets from other sources of healthcare data, including our institution's electronic medical record system. The ultimate goal for these tools is to allow clinical investigators to propose ideas worth studying and then "go fishing" in the database to see if these ideas are well represented in the clinical data at their institution.

Acknowledgments We would like to thank the leadership within the Department of Diagnostic Radiology at Henry Ford Hospital for their support as well as Dr Donald Peck and his team within the Department of Physics. We would also like to thank Dr Joseph Craig and Dr Courtney Scher within the Department of Musculoskeletal Radiology at Henry Ford Hospital for their contribution to the research project titled "Retrospective Review of Bone Marrow Signal Alterations Involving the Subchondral Bone Plate on Magnetic Resonance Imaging after Acute Knee Trauma" which served as the use case scenario for this project. Finally, we would like to thank Taryn Simon and Ward Detwiler at the Henry Ford Innovation Institute for their continued advisement of the Datafish Project during its ongoing research and development.

References

1. Koh H, Tan G: Data mining applications in healthcare. J Healthc Inf Manag 19:64–72, 2011
2. Asur S, Huberman B: Predicting the future with social media. Web Intelligence and Intelligent Agent Technology (WI-IAT) 2010 IEEE/WIC/ACM International Conference, Vol 1. 2010, IEEE

3. Singer P, Helic D, Hotho A, et al: What is Twitter, a social network or a news media?. 24th International World Wide Web Conference, Florence, Italy, 2015, (Best Paper Award)
4. Cerami E, Gao J, Dogrusoz U, et al: The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2(5):401–404, 2012
5. Zhang CQ, Petry D, Garzon JI, et al: PrePPI: a structure-informed database of protein–protein interactions. Nucleic Acids Res 41: D828–D833, 2012
6. Li P, Huang C, Fu Y, et al: Large-scale exploration and analysis of drug combinations. Bioinformatics 31:2007–2016, 2015
7. Demir E, Cary M, Paley S, et al: The BioPAX community standard for pathway data sharing. Nat Biotechnol 28:935–942, 2010
8. Cheng W, Rolls ET, Huaguang G, et al: Autism: reduced connectivity between cortical areas involved in face expression, theory of mind, and the sense of self. Brain 138:1382–1393, 2015
9. Radiological Society of North America: RSNA Informatics Radiology Reporting Initiative. http://www.radreport.org/, 2014. Accessed 10 Dec 2014