Focused Decision Support: a Data Mining Tool to Query the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Dataset and Guide Screening Management for the Individual Patient

Arjun Sharma¹ · Jason Hostetter² · James Morrison³ · Kenneth Wang⁴ · Eliot Siegel⁴

Published online: 18 September 2015
© Society for Imaging Informatics in Medicine 2015

Abstract The Prostate, Lung, Colorectal, and Ovarian Cancer (PLCO) Screening Trial enrolled ~155,000 participants to determine whether certain screening exams reduced mortality from prostate, lung, colorectal, and ovarian cancer. Repurposing the data provides an unparalleled resource for matching patients with the outcomes of demographically or diagnostically comparable patients. A web-based application was developed to query this subset of patient information against a given patient's demographics and risk factors. Analysis of the matched data yields outcome information which can then be used to guide management decisions and imaging software. Prognostic information is also estimated via the proportion of matched patients that progress to cancer. The US Preventative Services Task Force provides screening recommendations for cancers of the breast, colorectal tract, and lungs. There is wide variability in adherence of clinicians to these guidelines and others published by the Fleischner Society and various cancer organizations. Data mining the PLCO dataset for clinical decision support can optimize the use of limited healthcare resources, focusing screening on patients for whom the benefit to risk ratio is the greatest and most efficacious. A data driven, personalized approach to cancer screening maximizes the economic and clinical efficacy

and enables early identification of patients in which the course of disease can be improved. Our dynamic decision support system utilizes a subset of the PLCO dataset as a reference model to determine imaging and testing appropriateness while offering prognostic information for various cancers.

Keywords Decision Support · Data mining · Decision support techniques · Web technology

Background

Although cancer rates have steadily declined in the past two decades [1], the disease continues to represent a major source of morbidity and mortality across the globe. In 2010, cancer represented the second most common cause of death in the USA, with the average person having a 43.3 % lifetime risk of developing cancer [2]. There has been particular interest in cancer screening at the patient and population level, weighing the benefits of capturing cancers at an early stage with the financial costs and morbidity associated with screening. A number of recent controversies related to screening for breast, prostate, and thyroid cancers [3–5] highlight a growing need for clinical decision support to help assess a patient's individual risk and thereby stratify that patient into an appropriate screening paradigm.

Clinical decision support historically has often come in the form of expert consensus statements from professional or governmental organizations. The United States Preventative Services Task Force provides guidelines for and against the screening of a number of different cancer types based on the best evidence available [6]. Additionally, guidelines exist for the related problem of incidental findings and their appropriate management, including those described by the Fleischner Society for pulmonary nodules [7, 8] and those from the

✉ Arjun Sharma
arjunsharma33@gmail.com

¹ Department of Radiology, University of Maryland Medical Center, 22 S. Greene St, Baltimore, MD 21201, USA

² University of Maryland Medical Center, Baltimore, MD, USA

³ Oregon Hospital and Science University, Portland, OR, USA

⁴ University of Maryland Medical Center, Baltimore VA Medical Center, Baltimore, MD, USA

Society of Radiologists in Ultrasound (SRU) regarding asymptomatic ovarian lesions [9].

For the sake of simplicity, such decision support mechanisms typically use a small number of specified input parameters to assign patients to categories with little regard to additional specific/personalized factors which may influence the individual patient's cancer risk profile. It is left to the clinician to consider these factors to decide if a patient's particular cancer risk merits deviation from the management guidelines. For example, the Fleischner criteria accept as inputs nodule size and a subjective cancer risk while the SRU guidelines use imaging characteristics and menopausal status. In neither case would a family history of cancer or the patient's ethnicity impact the estimate of likelihood of disease or recommended follow-up. The ease of use of these guidelines allows for widespread adoption but limits the clinician's ability to provide a patient with his/her individual cancer risk and determine optimal follow-up and treatment tailored to that risk.

The rise of "big data" techniques, in concordance with the recent publication of massive clinical datasets, allows the development of data mining tools to more precisely target such recommendations. In particular, the Prostate, Lung, Colorectal, and Ovarian Cancer Trial (PLCO) provides a source of data unprecedented in scope and size for studying population-based cancer screening [10]. Repurposing the vast, multi-parametric clinical data gathered in the PLCO trial, we developed a decision support tool that allows clinicians to specify a number of demographic parameters in order to determine the risk of developing cancer in a given patient. The application further allows comparison of cohorts, such that a clinician may determine the incremental risk implicated by a specific parameter or multiple parameters. The absolute and relative risk of developing various cancers could then be used in real-time to help clinicians to tailor baseline cancer screening or follow-up for incidental findings that may develop into cancer.

Methods

The PLCO dataset was obtained via the online Cancer Data Access System (CDAS), administered by the National Cancer Institute (NCI) at the National Institute of Health (NIH). Access

to the dataset was made possible through an agreement signed by the authors and NCI, allowing use of the data in the development of a clinical decision support application as described here. Protected health information was not provided or needed for our study in accordance with the Health Information Portability and Accountability Act (HIPAA), and researchers are restricted from attempts to identify participants. The NCI retains ownership of the data, and an acknowledgement is required in any publication resulting from its use.

Initiated in 1993, the PLCO trial randomized 154,901 patients to two arms: an intervention cohort who would receive screening for cancers of the prostate (digital rectal examinations and blood test for prostate-specific antigen), lung (chest radiographs), colorectal system (flexible sigmoidoscopy), and ovaries (transvaginal ultrasound and blood test for the CA-125 tumor marker) and a control arm who would receive the standard cancer screening paradigm (see Fig. 1). A large volume of data regarding the patients' health status, risk factors, and dietary habits was collected at the trial outset by questionnaire. Outcomes assessed included the results of screening examinations, the development of cancer, the type, location, staging, and treatment of cancer, and mortality. The trial was initially designed to follow patients for at least 13 years, but data on patients will continue to be collected through 2015. A number of publications have explored various questions using the trial data, but to our knowledge, none have yet attempted to develop a decision support tool making the data readily accessible to the clinician.

A custom application was developed at the author's institution using a web-based interface to query a server-side database. Much of the technical framework used to the design the application has been previously described and adapted for use with the PLCO dataset [11]. A master table maps demographic data and baseline questionnaire answers to cancer outcomes. Additional tables are cross-referenced using a relational database schema for information regarding diet and screening examination findings. A subset of the data contained in the master table was chosen to be made available in the user interface in the initial version of the application: age, gender, race, height, weight, BMI, education, marital status, occupational status, family history of cancer, personal medical history, and variables regarding tobacco and alcohol

Fig. 1 The initial design of the PLCO trial

Fig. 2 A partial view of the tool with sample search parameters selected on the left and associated outputs on the right. The ability to add cohorts is also present on the upper right

use. Data from additional tables further provide access to specification of variables related to screening modalities.

A subscription-based Internet service was used to host a cloud-based web-server. As described by Morrison, the back-end application architecture was developed using Node.js, MySQL, and NGINX. The interface leverages several client-side user interface development libraries, including Bootstrap, AngularJS, and jQuery.

The application presents the client with an array of sliders and radio buttons with which a user may select a cohort of patients according to the variables described above. A query can be run with as few as one parameter selected or with parameters selected for every variable. If no variables are modified, the application returns the rate with which cancers were found in the trial, arranged by organ of origin in order of decreasing frequency. If a set of variables are selected, an “experimental cohort” is created. The number of patients meeting the criteria is returned, as well as the absolute rate of developing cancer, the odds ratio as compared to the overall trial group, and the associated statistical significance of that odds ratio (Fig. 2). Again, these numbers are categorized by cancer organ. For gender-specific cancers including prostate, penile, testicular, ovarian, uterine, cervical, and others, a

special function was incorporated to ensure chosen patients would only be compared against other patients of that gender.

A secondary function was developed allowing the comparison of multiple experimental cohorts. By allowing the user to compare two chosen groups, the ability to test the incremental contribution of individual parameters is made possible. The user selects a number of variables and assigns these values to “experimental cohort 1.” Changing perhaps just one variable, a second set can be chosen and assigned “experimental cohort 2.” The application can then compare the groups and return the comparative cancer risk. For example, one could create a query comparing current and former smokers, allowing the clinician to have concrete data in counseling a patient as to how smoking cessation will change her particular chances of developing cancer.

Results

Using the most recently published PLCO tables, the tool is capable of querying a total of 141,444 patients from the study. Patient ages ranged from 49 to 78 years of age. Height varied from 48 to 84 in. and weight from 70 to 399 lbs. With no

parameters selected, overall cancer rates were returned for the most common malignancies: prostate (11.2 %), breast (3.0 %), lung (2.3 %), and colorectal (1.5 %). Other malignancies with a nearly 1 % prevalence included melanoma (1.0 %), bladder (1.0 %), and non-Hodgkin's Lymphoma (0.8 %).

Males had a statistically significant increased chance of developing almost every cancer subtype as compared to females, with exception to thyroid (OR 0.58; 0.44–0.76 95 % confidence interval), peritoneal (0.14; 0.07–0.29), and breast (0.01; 0.01–0.01). Males are at a particularly high comparative risk for laryngeal (9.47; 5.9–15.21), esophageal (7.07; 4.88–10.25), and bladder (4.49; 3.91–5.15) cancers.

A history of previous or current tobacco use portended an increased risk of most cancer subtypes but was interestingly associated with a decreased risk for cancers of the prostate (0.81), breast (0.72), non-Hodgkin's lymphoma (0.88), and endometrium (0.84). This association may be due to a number of reasons; for example, smokers tend to live shorter lives and may not live long enough to develop sufficient prostate cancer risk; and smokers tend to be male, which would confound the effect of smoking on breast cancer as males have a much lower chance of developing breast cancer. Indeed, use of the cohort comparison functionality shows that female tobacco users have a statistically significant increased risk (OR 1.3) for breast cancer as compared to non-users. This is an example of the use of the cohort comparison feature to “control” for gender.

Although many more results can be drawn from use of the described tool by adjusting and combining all variables, such analysis is outside the scope of the present demonstration.

Discussion

Since 1993, 321 studies have been published using the PLCO dataset. Many of these have published “corner cases” for the data, examining, for instance, the relationship between vitamin D use and development of cancer, or obesity and its relationship to lung cancer.

The confluence of the rise of “big data” and the publication of several large trials allows for a new age of precision in personalized, data-driven clinical decision support. The immediate challenge is to make such data available to clinicians outside of research agreements and to present them the information in a simple and readily consumable format. The previously described NLST tool is one such solution but is limited in scope to lung cancer. The USPSTF has an online tool to guide screening recommendations [6] but accepts very few input parameters and returns are based on a limited set of algorithms.

We developed a tool that makes the vast data contained in the PLCO trial much more discoverable and allows a clinician to ask these questions in real-time by interacting with the user

interface or potentially automatically having these data extracted from the EMR. Our tool is hosted on the web and can be accessed from any web-enabled device. In a matter of seconds, it allows clinicians to establish the baseline rate for most types of cancer, the degree to which environmental and demographic risk factors are associated with development of cancers, the overall cancer risk profile for an individual patient, and the implication each risk factor has for that patient's cancer risk profile. We note that permission was given by the NCI to demonstrate the technology required to develop this tool and test its performance and utility; permission has not yet been granted to make this tool available to healthcare providers. Ultimately, the tool would ideally allow clinicians to personalize screening for individual patients, large insurers, and organizations to justify and reimburse deviations from standard screening guidelines and perhaps by patients to identify and modify the factors that influence their particular risks for different types of cancer.

Established screening recommendations exist for cancers of the breasts, colorectal tract, and most recently lungs [6]. Additional controversial screening modalities include the use of PSA for prostate cancer and ultrasonography or serum CA-125 for ovarian cancer. There is wide variability in adherence of clinicians to these guidelines and others published by the Fleischner Society and various cancer organizations [12, 13]. Data mining the PLCO dataset for clinical decision support can optimize the use of limited healthcare resources, focusing screening on patients for whom the benefit to risk ratio is the greatest and screening would be most efficacious. Additionally, the PLCO data could be utilized with the NLST dataset to establish a pre-test probability for lung cancer and potentially improve the predictive value of the NLST patient and imaging data for an individual patient. It could similarly be utilized as part of a Bayesian predictive model to improve lung CAD performance.

A number of factors were considered in the design of the application. Inclusion of all variables recorded by the PLCO tool, including baseline, dietary, and screening examination results, would make for an unwieldy and unusable interface. A choice was made to include only those variables which were felt to represent the widest interest possible: smoking history was included but thiamine supplementation was not. This design decision sacrifices the number of applications for the sake of usability. The next version of the software will allow clustering of similar parameters and easy selection of specific variables from the entire list for users. The increased specificity of additional variables also comes at the cost of power as more parameters are chosen, cohorts become smaller and more limited in their ability to find statistically significant results. By its nature, the tool is further limited by its source data and can only be generalized to the extent that its ~150,000 participants represent the overall population. To this end, biases such as the volunteer effect

imply that overall cancer and comorbidity rates are likely underestimated by the tool.

There are several directions in which we would like to advance the project. Automated data extraction from the electronic health record or even integration into the EHR would give clinicians easy access to a patient's overall cancer risk profile. On the output side, correlation of the malignancy rates and odds ratios for various cancers could be mapped to the values used by other white papers in their recommendations for screening, follow-up, and management of incidental findings. In addition to suggesting a screening paradigm, we would like to see the numbers generated by the tool incorporated as a pre-test probability into the screening examination itself and thereby help a radiologist assess the risk of an abnormal finding. Of course, once an examination is performed, those results should ideally be incorporated back into the tool to update that patient's cancer risk profile.

Ultimately, a robust system for cross-referencing our tool with other datasets, such as the NLST, the VA system, the NHS system, and any other similarly structured data would continue to add power and generalizability to the results returned. This is the hope as efforts to interoperate, automatically extract from the EMR, and apply big data techniques to healthcare data proliferate [14, 15]. To this end, the development of further clinical decision support tools depends fundamentally on continued efforts to structure healthcare data in all of its forms. For our part, we plan to explore additional options such as utilizing high performance computing systems to create an on-the-fly model of the selected subset of parameters and performing a statistical analysis such as a Jackknife regression or machine learning system such as a form of neural network.

Conclusion

The Prostate, Lung, Colorectal, and Ovarian Cancer (PLCO) Screening Trial dataset provides an unparalleled resource for matching patients with the outcomes of demographically or clinically comparable patients. Using these matched data, an individualized diagnostic decision support system can personalize imaging, testing, follow-up intervals, intervention, and prognosis. Ultimately, they can also be incorporated into CAD algorithms to improve diagnostic efficacy by providing a priori likelihood of disease information. A data-driven, personalized approach to cancer screening maximizes its economic and clinical efficacy and enables early identification of patients in which the course of disease can be improved. The use of these large research and ultimately clinical

databases combined with high performance computing will undoubtedly have a major impact on the practice of medicine.

References

1. Edwards BK, Noone AM, Mariotto AB, et al: Annual Report to the Nation on the status of cancer, 1975–2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer* 120(9):1290–1314, 2014
2. Murphy SL, Xu J, Kochanek KD: Deaths: final data for 2010. *National Vital Statistics Report* 61:4, 2013
3. Suzuki A, Ishida T, Ohuchi N: Controversies in breast cancer screening for women aged 40–49 years. *Jpn J Clin Oncol*, 2014. doi:10.1093/jco/hyu054. **online only**
4. Barry M: Screening for prostate cancer – the controversy that refuses to die. *N Engl J Med* 360:1351–1354, 2009
5. Ahn HS, Kim HJ, Welch HG: Korea's Thyroid-Cancer “Epidemic” – Screening and Overdiagnosis. *N Engl J Med* 371:1765–1767, 2014
6. MacMahon H, Austin JH, Gamsu G, Herold CJ, Jett JR, Naidich DP, et al: Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. *Radiology* 237(2):395–400, 2005
7. U.S. Preventative Services Task Force (various years). Recommendations for adults. Retrieved from <http://www.uspreventiveservicestaskforce.org/adultrec.htm>
8. Naidich DP, Bankier AA, MacMahon H, Schaefer-Prokop CM, Pistolesi M, Goo JM, et al: Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. *Radiology* 266(1):304–317, 2013
9. Levine D, Brown DL, Andreotti RF, Benacerraf B, Benson CB, Brewster WR, et al: Management of asymptomatic ovarian and other adnexal cysts imaged at US. *Radiology* 256(3):943–954, 2010
10. Gohagan JK, Prorok PC, Hayes RB, Kramer BS: The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the National Cancer Institute: history, organization, and status. *Control Clin Trials* 32(6):251S–272S, 2000
11. Morrison JJ, Hostetter J, Wang K, et al: Data-driven decision support for radiologists: re-using the national lung screening trial dataset for pulmonary nodule management. *J Digit Imaging* 28(1): 18–23, 2014
12. Lacson R, Prevedello LM, Andriole KP, Gill R, Lenoci-Edwards J, Roy C, et al: Factors associated with radiologists' adherence to Fleischner Society guidelines for management of pulmonary nodules. *J Am Coll Radiol* 9(7):468–473, 2012
13. Eisenberg RL: Ways to improve radiologists' adherence to fleischner society guidelines for management of pulmonary nodules. *J Am Coll Radiol* 10(6):439–441, 2013
14. Frankovich JF, Longhurst CA, Sutherland SM: Evidence-based medicine in the EMR era. *N Engl J Med* 365(19):1758–1759, 2011
15. Kim W, Piraino D: Big data in healthcare: myth, hype, and hope – a point/counterpoint. Town hall meeting conducted at the meeting of the Society of Imaging Informatics in Medicine. National Harbor, MD, 2015