Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computational Advances in Cancer Informatics (B)

Design and Implementation of a Comprehensive Web-based Survey for Ovarian Cancer Survivorship with an Analysis of Prediagnosis Symptoms via Text Mining

Jiayang Sun[1], Kath M. Bogie[2,6], Joe Teagno[3], Yu-Hsiang (Sam) Sun[4], Rebecca R. Carter[1], Licong Cui[5] and Guo-Qiang Zhang[5]

[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA. [2]Departments of Orthopaedics and Biomedical Engineering, Case Western Reserve University, Cleveland, Ohio, USA. [3]KeyBank of Cleveland, 127 Public Square, Cleveland, Ohio, USA. [4]Darfon Electronics Corp., Taoyuan County, Taiwan. [5]Department of Electrical Engineering and Computer Science, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio, USA. [6]Louis Stokes Cleveland Department of Veterans Affairs Medical Center (LSCDVAMC), Cleveland, Ohio, USA.

ABSTRACT: Ovarian cancer (OvCa) is the most lethal gynecologic disease in the United States, with an overall 5-year survival rate of 44.5%, about half of the 89.2% for all breast cancer patients. To identify factors that possibly contribute to the long-term survivorship of women with OvCa, we conducted a comprehensive online Ovarian Cancer Survivorship Survey from 2009 to 2013. This paper presents the design and implementation of our survey, introduces its resulting data source, the OVA-CRADLE^{TM} (Clinical Research Analytics and Data Lifecycle Environment), and illustrates a sample application of the survey and data by an analysis of prediagnosis symptoms, using text mining and statistics. The OVA-CRADLE^{TM} is an application of our patented Physio-MIMI technology, facilitating Web-based access, online query and exploration of data. The prediagnostic symptoms and association of early-stage OvCa diagnosis with endometriosis provide potentially important indicators for future studies in this field.

KEYWORDS: survey, design, database, informatics, CRADLE^{TM} (Clinical Research Analytics and Data Lifecycle Environment), early diagnosis, ovarian cancer, endometriosis, prediagnosis symptoms, text mining, correspondence analysis, (CART) Classification and Regression Tree, and Random Forest

SUPPLEMENT: Computational Advances in Cancer Informatics (B)

CITATION: Sun et al. Design and Implementation of a Comprehensive Web-based Survey for Ovarian Cancer Survivorship with an Analysis of Prediagnosis Symptoms via Text Mining. *Cancer Informatics* 2014:13(S3) 113–123 doi: 10.4137/CIN.S14034.

RECEIVED: June 02, 2014. **RESUBMITTED:** December 21, 2014. **ACCEPTED FOR PUBLICATION:** December 25, 2014.

ACADEMIC EDITOR: J.T. Efird, Editor in chief

TYPE: Original Research

FUNDING: GQZ acknowledges support from National Institutes of Health (NIH) / National Center for Advancing Translational Sciences (NCATS) award UL1TR000439 and J. Sun from National Science Foundation (NSF) award DMS−0306202. The authors confirm that the funders had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

ETHICS: The study was reviewed and found to be exempt from IRB review under 45CFR.46 because it involved the use of de-identified, survey procedures only.

JOURNAL OF CANCER INFORMATICS: Invited submission by JCI.

CORRESPONDENCE: jsun@case.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE). Published by Libertas Academica. Learn more about this journal.

Introduction

Ovarian cancer (OvCa) is the fifth leading cause of cancer deaths in women and the most lethal gynecologic malignancy, with more women dying from this disease than all other gynecological cancers combined. The overall 5-year survival rate for all OvCa patients is 44.6% and is markedly lower at 27% or less for women diagnosed at an advanced stage.[1] According to the American Cancer Society, in 2014 alone there will be an estimated 21,980 newly diagnosed cases and 14,270 deaths in the United States (US).[2] The lifetime risk for women in the US to develop OvCa is 1 in 73.[3] The most recent cancer incidence report by Surveillance, Epidemiology, and End Results (SEER) program indicates that there were 188,867 Americans living with OvCa in 2011,[1] with some as young as 9 years of age, and others over the age of 84. The median age at diagnosis for OvCa is 63.[1,4] These statistics indicate not only that all women are impacted but also that the mortality rate for OvCa is unusually high.

OvCa has been termed the "silent killer" because many women do not show clear symptoms at an early stage, and a large percentage of the general public and health professionals lack awareness of early signs of OvCa. For example, a recent study[5] showed severe knowledge gaps even in OvCa awareness (and hence about its early signs) among the general population. Other studies[6,7] have presented gaps in care by health professionals. When OvCa is diagnosed at a late stage, effective treatments are limited. The US national OvCa care expenditure was $5.12B in 2011, making it one of the most expensive cancers to treat.[8] However, more than 80% of patients do have symptoms, even when the disease is still limited to the ovaries.[9] In addition, some late-stage patients do live a long time. Of the 188,867 Americans living with OvCa, 15,675 had lived at least over 36 years since diagnosis.[4] The clinical and social challenges impacting OvCa management motivated us to carry out a comprehensive Web-based survey of OvCa patients. Our goal was to assess multiple aspects of OvCa management, initial symptoms that led to diagnosis, medical history, possible genetic and environmental factors, cancer treatments, and lifestyle and behavior before and after diagnosis. All these factors may affect survivorship.

This paper presents the design and implementation methodology of the Web survey, provides a glimpse of the database and functions of the OVA-CRADLE^{TM} (Clinical Research Analytics and Data Lifecycle Environment), and illustrates an analysis of the prediagnosis symptoms, one of the 1502 fields in our data, using novel text mining and statistic techniques. The analysis of prediagnosis symptoms is interesting in its own right. It also reveals evidence that could be used for narrowing the knowledge gap in OvCa and provides a basis for future studies about early OvCa diagnosis. This paper also demonstrates how a cancer informatics project from A to Z can be done effectively using modern technology with an interdisciplinary team. This is the first of a sequence of papers expected from this cancer informatics project.

Methods

This section provides our methods for each component of the OvCa project in its logical development order: Design and Implementation of the Web survey, Recruitment and Data Collection, Development of data resource OVA-CRADLE^{TM}, Text Mining of free-text responses to prediagnosis symptoms as a special example of data analysis that can be done with our data, and Statistical Analysis Methodology of the numerical data from text mining. The results of our analysis are presented in the Results section and the conclusions are provided in the last section.

Survey design. The survey objective is to determine factors that contribute significantly, albeit partially, to the long-term survivorship of OvCa survivors, using information directly from patients and caregivers. Since cancer development is a multiphased and multifactorial process, a good survey needs to be *comprehensive*. The survey questions also needs to be *sensitive* to the feelings of patients who respond to the survey. Next, the survey needs to reach as many people as possible, with resulting data to be of high *quality* and in a format that is easy to manage. Therefore, our design was *Web-based* and included a *quality control* scheme and a connection to a state-of-the-art *database*, which allowed instant and automatic data capture, safe and secure data storage, and a fast archive of entire data on demand to a standard data format. The standard data format is required for easy exporting data to downstream informatics platforms, such as OVA-CRADLE^{TM}, and to standard statistical software systems for data analysis and for text mining.

With the above design objectives, initial development of the Web-based survey started in May 2009. The survey design and implementation were completed in collaboration with both local and national stakeholders, including OvCa survivors, support groups, scientists, and clinical professionals. It was launched via Case Western Reserve University in September 2009, with links to the survey portal published at the Women Cancer Network, wcn.org, from 2009 to 2013.

Comprehensiveness. The survey encompassed 1502 fields grouped into 15 sections encompassing all stages of OvCa management from the patient's perspective. This is the most comprehensive survey on women diagnosed with OvCa. Factors assessed in the survey included physical, physiological, environmental, behavioral, and clinical care factors in OvCa, as well as symptoms that led to diagnosis, medical history, treatments and pre- and postdiagnosis lifestyle. For the complete survey, see the Ovarian Cancer Survivorship Survey archive.[10]

Sensitivity. The survey was reviewed by patient stakeholders and tested by additional collaborators before its public launch. The questions were worded to be sensitive to a patient's emotional status, to minimize any possible anchoring effect, and to conform to standards set by PROMIS[11] and FACIT tools[12] for most questions when applicable. The environmental factor was coded on a scale of 1–6, based on a map of excess cancer risk due to air pollution, which appeared in *USA Today*[13] and was generated using data compiled by the Environmental Protection Agency (EPA).[14] We programed our survey to allow an automatic insertion of the scale by means of a drop-down menu (I, II, III, IV, V, VI) for the geographic region to quantify the environmental factor where the participant had lived for a significant period of time prior to diagnosis. In this way, the participants did not need to be concerned with their exact geographic location being recorded in our database.

Accessibility. The survey was completed on an anonymous, voluntary basis through a Web page. Therefore the study was single-blind. Each participant could choose her own login name and password. This option permitted unlimited revision of survey entries by participants before the end of the study. In order to increase outreach to a broad audience, study respondents could either submit their responses online or mail a hard copy to the study host without a mailing address. All data were de-identified at collection.

Quality control and support. An automatic alert was implemented to send a warning if an input was outside a reasonable range of the possible values of a particular variable. If the value was indeed the intended input, the value was still permissible for discovery of unexpected evidence. Support for a standard source of information was provided, such as those for complementary alternative medicine (CAM) and environmental coding. Since the recruitment was conducted via credible OvCa networks and testers were given comprehensive instructions, the chance of having an unwelcome intruder who would answer consistently for many/most of 1502 questions was small. The inconsistency and outlier detected by statistical techniques provided additional quality control for our subsequent data analysis.

Risk minimization. The Web-based anonymous survey of OvCa survivors and their caregivers, followed by analysis of de-identified data was considered minimal risk and a waiver of informed consent was obtained from the local Institutional Review Board.

Survey implementation. *Our backend database and management system* were developed and implemented using Hypertext Preprocessor (PHP), a common language for Web development that can easily interact with an open-source MySQL database management system. Our implementation in PHP with MySQL as the backend also allowed the entire de-identified data to be exported in comma-separated values (CSV) format, a standard format for archived data, on-the-fly from the Web server, by those with administrative privilege. Our implementation facilitated continuous monitoring and interim analysis of incoming data, as well as export to other systems for data manipulation, exploration, statistical analyses, and text mining.

Recruitment and data collection. All OvCa survivors were invited to participate in the study after the implementation of the Web survey. Non-English-speaking survivors were excluded if they could not complete the survey in English.

Study awareness was achieved via multiple venues. A call for participation was distributed via oncologists interested in publicizing the survey, OvCa support groups, websites, and community newsletters including the Foundation for Women Cancer, also called Women Cancer Network (wcg.org) and Gynecological Cancer Foundation (www.thegcf.org), Conversations (the international newsletter for those who fight OvCa), OROC (or OutRun Ovarian Cancer, www.oroc.org, a volunteer-driven, nonprofit organization dedicated to raising OvCa awareness and money for research and education), and the Gathering Place (a caring community for those touched by cancer, http://www.touchedbycancer.org). The launch effort was completed in mid-2010 but the survey remained open and available online until the end of 2013 at the WCN. Continued survey awareness has been maintained by word-of-mouth and references to the study at Survivorship Courses by the WCN.

Despite no active recruitment effort since early 2011, the response rate has remained steady (Fig. 1), as OvCa survivors, especially newly diagnosed patients, visit the Website of the Foundation for Women Cancer. A surge in responses during the months of May was most likely due to the annual wave of presentations of OvCa Survivorship courses, sponsored by the WCN. Another surge in December 2012 was probably due to the presentations by Dr Sun when she gave talks about this work during her research visits.

OVA-CRADLE™ for data query and exploration. A *research resource* OVA-CRADLE was built *using our Web data*. The collected data as of 03/24/2013 had 943 unique patients (after removing 9 known test cases from 952 entries), each with 1502 fields. Given data in CSV format directly from our Web survey, they were ported to OVA-CRADLE™ for Web-based access and online query and exploration. Apparent outliers can be easily identified by data exploration at OVA-CRADLE™.

OVA-CRADLE™ is adapted from Physio-MIMI, a patented data integration environment[15] funded as one of the three informatics pilot projects under CTSA RFP 08–001.[16] Physio-MIMI uses ontology for directly driving the federated query interface VISAGE (VISual Aggregator and Explorer)[17] and for integrating autonomous data resources through the database to ontology mapper called Physio-Map.[18] The Physio-MIMI architecture makes use of ontology beyond its traditional role for terminology standardization, resulting in a flexible framework with domain ontology as a "plug-and-play" component capable of harmonizing disparate data sources without requiring a uniform data model. Because of the flexibility and plug-and-play system architecture, Physio-MIMI has been adapted to a variety of settings, such as for multicenter epilepsy study[18–20] and for querying Medicare and Medicaid data.[21]

OVA-CRADLE™ makes the data both searchable and queryable through its VISAGE interface. The VISAGE interface uses individual query widgets, automatically generated from a study data dictionary or a proper ontology, to define a set of cohort query constraints. Users can define logical connectives, such as AND, NOT, OR, between each query widget to build a complex query for data exploration, or simply for facilitating analysis of a subset of our data. The results of the queries are displayed in VISAGE for easy browsing and graphical rendering. The resulting comprehensive data and database can be used to investigate many questions about OvCa, test a hypothesis of interest, and generate hypotheses for future studies.

For example, Figure 2 shows an example of the VISAGE query builder interface for OVA-CRADLE™. Illustrated is a very simple sample query of blood type A OvCa patients (193) with three or four recurrences (39). This query resulted in a total of five patients. Figure 3 illustrates the on-the-fly

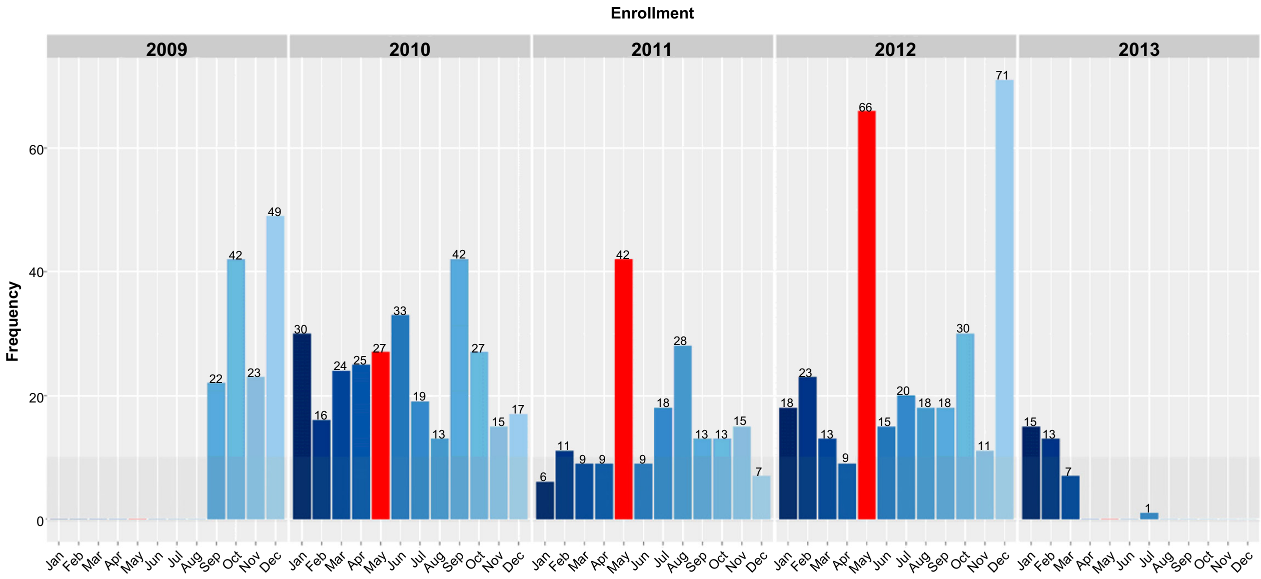


Figure 1. Participants response distribution (September 2009 to March 2013).

scatter plot, one of the built-in tools of VISAGE. Within a base query, additional responses from patients, such as stage and blood type, can be rendered on the same Web-based query interface, classified by different combinations of the stage and blood type for easy comparison of their distributions across these combinations. It can be seen from Figure 3 that the blood type A is a dominant type in this OvCa population.

Text mining strategy. We present methodology in this section for text mining prediagnosis symptoms. From among the 1502 available fields, we focused on *one* field for a key question of interest to us: investigation of possible patterns in prediagnosis symptoms. Prediagnosis symptoms are best studied from patient-reported outcomes. To minimize recall bias, we designed a free-text response question in our survey so that there would be no reminders, hints, anchoring, or other biasing effects in the data collected. Our specific question was

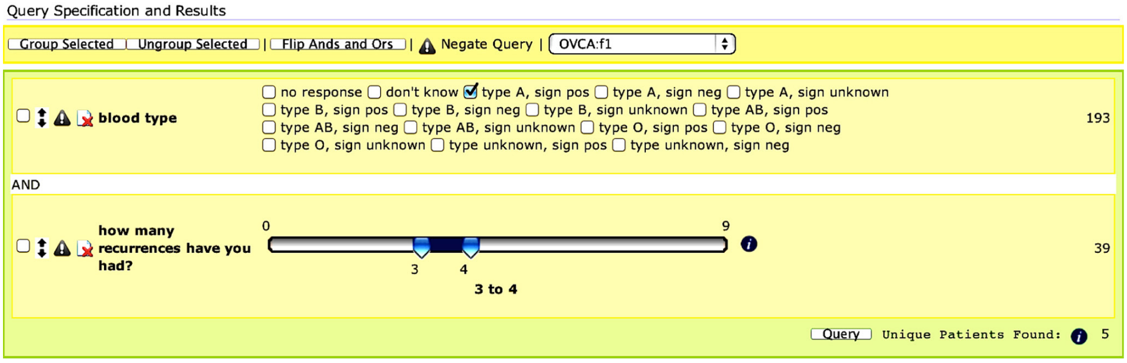
"How did you or your doctor discover that you had ovarian cancer?"

Our instruction for answering this question was

"In your free-response here, please make sure to include symptoms *that led you to the doctor and the tests* you had *before* the confirmation from the pathology report."

In this free-text response, women were prompted to include all symptoms and tests that led to their cancer diagnosis. Although a 100% elimination of possible recall bias can never be guaranteed, our free-text strategy for collection of prediagnosis symptoms is a major improvement to a pure multiple-choice strategy and it facilitates findings not known before.

It should be noted that text mining of these free text responses provides information on both the prediagnosis symptoms, reported here, and the diagnostic path. The impact of medical practices on the pathway to diagnosis to the diagnostic stage is beyond the scope of the current paper and will be addressed in a future paper. In order to trace the diagnostic path of each of these respondents, we adopted a modified version of the vector space model (VSM) to support our text-mining techniques.



Query Specification and Results

Group Selected | Ungroup Selected | Flip Ands and Ors | Negate Query | OVCA:f1

□ blood type

□ no response □ don't know ☑ type A, sign pos □ type A, sign neg □ type A, sign unknown
□ type B, sign pos □ type B, sign neg □ type B, sign unknown □ type AB, sign pos
□ type AB, sign neg □ type AB, sign unknown □ type O, sign pos □ type O, sign neg
□ type O, sign unknown □ type unknown, sign pos □ type unknown, sign neg

193

AND

□ how many recurrences have you had?

0 9
3 4
3 to 4

39

Query Unique Patients Found: 5

Figure 2. The VISAGE query builder interface for OVA-CRADLE^{TM}. Illustrated is a simple sample query of blood type A OvCa patients (193) with three or four recurrences (39). This query resulted in a total of five patients.

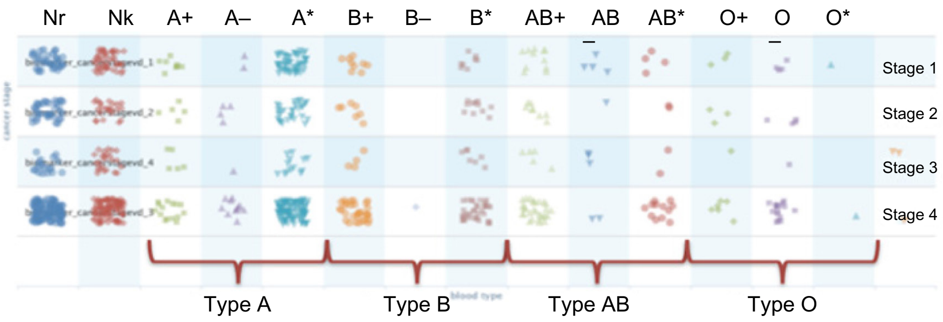


Figure 3. One of the built-in tools of VISAGE is the combination plot. Within a base query, additional responses from patients, such as staging and blood type, can be plotted on the same Web-based query interface.

Notes: Here the row labels indicate blood types as listed in Figure 2, with Nr = No response, Nk = Do not know, A+ = Blood type A with a positive sign, A– = Blood type A with a negative sign, A* = Blood type A, unknown sign, B+, B-, B*, AB+, AB-, AB*, O+, O-, and O* are defined similarly as those of A's, and Un = Type unknown, a sign is given (not interesting to show).

There are three steps for properly analyzing text data: (1) Exploratory Data Analysis (EDA) of the entire data to remove outliers and invalid data and to decide how to build text mining key elements; (2) Text mining to transform text into meaningful numerical data; and (3) Statistical analysis of the numerical data.

EDA. Reported symptoms that motivated women to seek clinical follow-up and subsequently led to diagnosis were categorized based on reported OvCa stage (Stage 1–4). Data were explored using statistical graphics and EDA both in R and with OVA-CRADLE$^{\mathrm{TM}}$. Of 943 survivors and their proxies, at least 905 were valid data entries. Invalid entries included our test cases, apparent outliers, and those with wrong or impossible data entries. Women who did not know their initial diagnostic stage were excluded from further analyses of prediagnosis symptoms, which led to a population of 872 women (= 905 – 33). All apparent outliers and wrong or impossible data entries (such as a reported weight of 4,233 pounds or cancer diagnosis age of 0 years old) were also excluded.

Text mining transformation. Of the 905 surveys, 535 contained a valid answer to the question: "How did you or your doctor discover that you had ovarian cancer?" 19 of the 535 respondents were unstaged, which leaves 516 (= 535 – 19) staged respondents with text. The validation was accomplished by two trained students who independently assessed all the text responses and also via their cross-examinations. In order to achieve our goal of tracing the prediagnosis symptoms of each of these 535 respondents, we needed a mechanism by which to identify and extract the prediagnosis symptoms and patterns of each subject.

We followed a modified version of the VSM,[22] which underlies newer text-mining techniques such as nonnegative tensor factorization[23] and kernel methods.[24] The key function of a VSM is judging the extent to which a single document matches a set of search criteria. Thus, the VSM has two requirements (a) that the document be indexed, ie, scanned for the presence of keywords of interest (*terms*); and (b) that the indexed terms be weighed against the query terms to determine whether the document matches. With these two requirements satisfied, the VSM model assigns one dimension for each indexed term, but it is optimized to perform well on the sparse, high-dimensional term-document matrices that tend to be generated from natural language text mining.

Based on medical knowledge, initial EDA of data, and later re-examination of data, we decided to use 21 terms, representing 21 groups of symptoms. To identify these specific 21 terms from our survey data, we used the tm()[25] package for GNU *R*. We treated each survey response as a separate document and examined the whole corpus for those terms that carried the greatest relevance and frequency. We expanded the list of terms to include additional topics that we wished to study, and for each of the 21 terms, we assembled a list of synonymous terms and allowed for typos in expressing the same symptoms. For example, the term for symptom 5 "chance," represents a set of words {accident, (annual|anual).(gyn|gync|ginecological), birth, cough, c-section, pap, routine|physical} and their "or" option by piping "|" and the linkage, which indicates a diagnosis alerted from either a CT scan following an *accident*, or actions from an *annual gynecological appointment* or *physical* [where *gynecological* may be mistyped as *ginecological* by a respondent], or discovery at *birth* of a child, and from a *routine visit* to a doctor's office. Here the dot "." between two groups of words requires that one word from each of two groups must both occur, such as "annual gyn" exam. The term for *symptom 21, breath and ascites,* shortened as "btheA," consists of {ascites|fluid, bad.cough, breath(e), shortness}.

Our application of the VSM model worked well with a small selected set of 21 group terms. With a relatively small number of dimensions, we were able to count this set of terms from each respondent and perform analysis using Unix scripting. Manual auditing of the text mining results was also performed by sampling the cases, as a quality control measure.

Each survey response was indexed into a term-document matrix storing the number of hits for each term by each document. This produced a numerical data matrix of responses, ready for formal statistical analyses described in the next section.

Statistical analysis methodology. Our analyses used a three-tiered methodology. Descriptive summary of the data and their comparison of demographics with national databases were used to validate our study population for prediagnosis symptoms. Rank-ordered symptoms per each of four disease stages provided an important view of the association of symptoms with each stage. Statistical analysis techniques, Correspondence analysis (CA), and Classification and Regression Tree (CART) analysis (with Random Forest [RF]) were our choices of nonparametric methodology for studying the association of symptoms with the disease stages from two other views. These techniques do not impose stringent parametric assumptions for the underlying distribution of the study data.

Results

Summary of survey data. The distribution of survey response was plotted in Figure 1, from September 2009 to March 2013. The Survey was open to all women who were recruited through multiple avenues in the first 18 months, linked at wcn.org. Among the first 18-month study population, nearly 17% women had *survived* for more than 5 years and over 12% for more than 10 years. Thus, the percentage of long-term survivors in our survey from this period is similar to that of the general population of OvCa survivors. Among the participants who responded to our study after 2011, there were more newly diagnosed patients participating in the survey as they visit wcn.org. The accounts from newly diagnosed patients on the symptoms that led them to the doctors and their tests before the confirmation from their pathology reports are more accurate than those who were diagnosed a long time ago. Using the shortest possible time interval between the prediagnosis symptoms (the event/illness of interest) and the study (the data collection time) is recommended as the first strategy by Feinerer et al.[25] and the second strategy by EDC[26] to prevent or minimize recall bias from patient-reported symptoms.

There were *missing values*, predominantly concerned with CAM and the foods in the participants' diet. Nevertheless, there remains extremely valuable preliminary information in many of the data fields collected. Our experience with this survey also provides valuable insights for designing our next phase study.

The demographics of 535 patients who provided text responses for prediagnosis symptoms are compared with the full data and those from national statistics. Table 1 shows the comparison in terms of the distribution of diagnostic stages. Our survey data (*N* = 872, excluding 33 unstaged from a total of 905) have a similar distribution to that of the subpopulation who provided text responses (*N* = 516, excluding 19 unstaged cases from 535). This subpopulation also has a similar distribution to that of the most recent SEER 18 data. The national data from earlier years have a much smaller percentage of women diagnosed at Stage 2 than SEER 18 and our data, both of which are more recent. Both data sets have more balanced rates for early stages than the older, national cancer data. More balanced data can provide more accuracy in an analysis of comparing different symptoms by different stages.

We also compared our data with national statistics in terms of age distribution. The distribution shapes of both the age at diagnosis and the age at our study time are comparable to those of the SEER's diagnostic[1] and prevalence data, respectively. The prevalence counts were computed by multiplying the SEER's age-specific prevalence rates (0.0002%, …, 0.1928%) in the NIC's report[4] with the age-specific women population counts given by the 2010 US census.[28] However, our population were slightly younger and had very few very young and very old participants, which is typical of Internet Web-based studies.

Comparison of rank-ordered frequencies. The results are shown in Table 2 and Figure 4. The most direct indication of OvCa is a *mass* in the abdominal region. Overall, this was the most common diagnosis symptom for women with Stage NA (not shown), Stage 1,2,4 OvCa, and the second common with Stage 3, while a *chance* encounter at the doctor moved into the top five symptoms for Stage 1 OvCa diagnosis and dropped a bit to the seventh top-ranked symptom for both Stage 2 and Stage 4, but dived below half of all symptoms for Stage 3, the most likely diagnosed stage. This is reasonable as doctors should be able to help diagnose OvCa easily when it is Stage 4, even if the patient only comes in for an annual exam,

Table 1. Comparison by diagnosis stages.

DATASET		CANCER STAGE			
		STAGE 1: N (%)	STAGE 2: N (%)	STAGE 3: N (%)	STAGE 4: N (%)
1998–2004 US National Cancer population database [34] (N = 79,568)		17,248 (21.7%)	6,642 (8.4%)	34,592 (43.5%)	21,086 (26.5%)
1993 SEER population [35](N = 23,186)		5,503 (24%)	819 (4%)	14,863 (72%)	
2004–2013 SEER 18 population[1] [1, 4]		16%	19.1%*[2]	64.9%	
Survivorship population	All (N = 872) Text (N = 516)	184 (21%) 101 (19.6%)	135 (15%) 82 (15.9%)	465 (53%) 281 (54.5%)	88 (10%) 52 (10.1%)

Notes: [1](16%, 19.1%, 64.9%) is obtained by reweighting (15% [Local], 18% [Regional], 61% [Distant], and 6% [Unstaged]) in [1] after removing the unstaged cases. [2]"Regional" by SEER is defined to "spread to regional lymph nodes"; thus, some of these 19.1% patients should be in Stage 3 based on the staging definition of OvCa at cancer.org. [3]*N* = 872 (excluding *n* = 33 unstaged from a total of 905) and *N* = 516 (excluding *n* = 19 unstaged from 535 with text responses).

Table 2. OvCa statistics of prediagnosis symptoms.

NAME (ABBR)	EATING (EAT)	PELVIS (PLV)	URINATION (URIN)	MASS (MASS)	CHANCE (CHNC)	ANEMIA (ANEM)	MENSTRUAL (MENS)	INDIGESTION (INDG)	FATIGUE (FATG)	WEIGHT (WGT)	CONSTIPATION
Stage1	40	27	20	63	24	4	42	14	13	13	10
Stage2	32	21	8	44	16	2	27	20	10	8	14
Stage3	150	64	44	118	44	2	79	79	56	46	59
Stage4	24	16	4	24	11	0	12	16	10	10	10
NAME (ABBR)	BACK (BACK)	INTER-COURSE (INT)	PAIN (PAIN)	FAMILY HISTORY (FAMH)	INFECTION (INFE)	ENDOMETRIOSIS (ENDO)	KIDNEY (KDNY)	CANCER (CA)	BREATH+ ASCITES (BTHEA)	OTHER (OTH)	UN-KNOWN
Stage1	8	4	17	13	6	13	1	1	6	6	
Stage2	6	1	22	9	4	5	1	2	8	9	
Stage3	23	10	60	30	10	9	8	4	39	23	
Stage4	3	2	10	6	4	0	1	2	22	10	

or at an ER, or by a CT scan following an car accident, say. The *chance* association with early-stage diagnosis more than with Stage 3 indicates the importance of care and response by a primary physician and a gynecologist. It supports the call for improved primary care for cancer survivors[29] and also for women in general, as the US population ages in the next two decades.

Compromised digestive functions, eg, eating and indigestion, were frequently reported as the initial diagnosis symptom *at all stages*. Eating problems were the most common diagnosis symptom for women with Stage 3, the stage most OvCa patients were diagnosed at, and tied with mass as the most common symptom for Stage 4. At all stages, another digestive function, constipation was not reported as frequently as these two digestive functions, though it did rank seventh for Stage 3.

Pelvic pain was among the top five most frequently reported symptoms. Nonspecific pain was most frequently reported by women with Stage 2 OvCa at diagnosis. Women with other stages of OvCa report nonspecific pain moderately frequently. Other symptoms that are considered to be generally diagnostic of cancer, eg, *fatigue and weight loss*, were present, but not reported with as high frequencies by women with OvCa as with other symptoms, though a bit more at Stage 4 diagnosis.

Breathing problem and/or ascites (btheA) were added as one group symptom, as it is associated particularly with Stage 4 diagnosis, with patients reporting a persistent or chronic "bad cough" or shortness of breath.

A relevant *family history* of breast, colon, or uterine cancer was a low, *stable*-frequency symptom for all stages, ranked 10th with Stage 1, 11th with Stage 2, and 13th with Stages 3 and 4. This is consistent with studies that determined: OvCa has some small association with breast cancer and perhaps colon cancer, but the vast majority of women diagnosed with OvCa do not have known high-risk genetic mutations.[30] A personal history of *endometriosis* was reported as a primary diagnostic symptom with rather different frequencies across the four stages, with the highest frequency at Stage 1 OvCa, lower at Stage 2, and much lower with Stage 3 and 4 OvCa.

Correspondence analysis. CA for categorical data arranged in contingency tables is analogous to Principal Component Analysis for continuous data. It is useful for studying the particular association of row and column variables – here diagnostic stages and 21 symptoms given in Table 2 – if the null hypothesis of no association between the row and column variables is rejected. In our case, the null hypothesis of no association between diagnostic stage and the symptoms was rejected using a chi-square test with a P-value < 0.000041. In a CA plot, such as our Figure 5, points that are away from the origin (0,0) but close to each other have strong positive associations, while points that are far away from each other or close to (0,0) do not have a specific association. The further away of two relatively close points from (0,0) the stronger their association is.

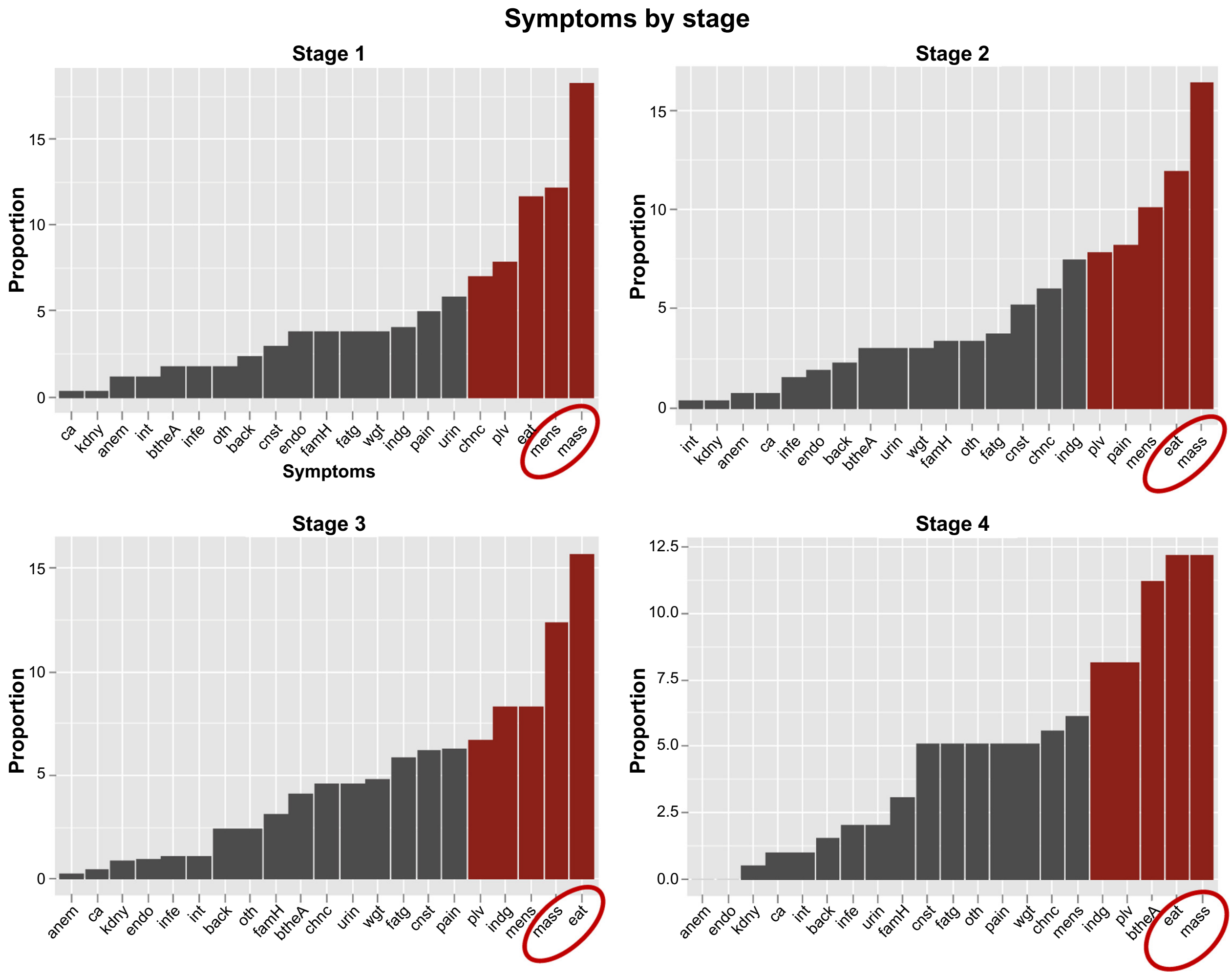


Figure 4. Ranked symptoms by stage. *Eating* and *mass* were tied for Stage 4. There are more chance findings in earlier stages. Endometriosis is significantly higher in Stage 1 than in other stages. See Table 2 for the definitions of the abbreviated symbols for prediagnosis symptoms.

Here the four circles in Figure 5 can be shrunk or expanded to indicate more focused or more relaxed groupings; and the points close to (0,0) should not be interpreted too much. Figure 5 shows that Stages 3 and 4 are on the left side of the graph while Stages 1 and 2 are on the right. Breath and ascites are strongly associated with Stage 4 diagnosis, while Stage 1 is closest to endometriosis. Although mass as a prediagnostic symptom is a top-ranked, nonspecific symptom for all stages (Fig. 4), it has a stronger association (Fig. 5) with Stages 1 and 2 than with Stages 3 and 4. This is reasonable as a person who is more aware of her own body before the medical diagnosis has a better chance for an earlier diagnosis. Stage 3 diagnosis is most strongly associated with eating and also strongly with many symptoms. Figure 4 presents relationships specific to each stage according to frequencies, while Figure 5 discriminates the differences among the four stages.

CART and Random Forest (RF) analysis. A woman can have multiple symptoms. A nonmodel-based flexible procedure for classifying diagnostic stages based on multiple co-existing symptoms is a tree-based model, such as the Classification and Regression Tree (CART) or the recursive partition tree, called rpart() in the open-source R. RF, denoted as randomForest() in R, grows many classification subtrees[31] and provides a classification based on the majority vote of these trees for each case but not one concrete tree-based model as CART and rpart() would. RF to us is more useful to rank the importance of variables, ie, symptoms in the classification of stages; it can also validate our final model from CART or rpart(). The CART analysis shown in Figure 6 was the *best tree* model obtained using CART and rpart() based on cross-validations and confirmation by RF. The tree demonstrated that the symptom *breath and ascites* was the most important predictor for late-stage classification and there was an association with endometriosis and a (self-diagnosed) mass in the abdomen for early-stage diagnosis, consistent to the finding in Figure 5. We annotated the CART by providing the probabilities of leading to each of the four-stage classifications, instead of only one outcome based on the highest probability as it would in a standard CART. If the node *other*, consisting of words {arm. pits|swollen, diarreah|diaria|diarrhea, nausea, puke, vomit} was pruned in Figure 6, ie, if we did not consider the presence or not to this particular *other* symptom, the classification rule

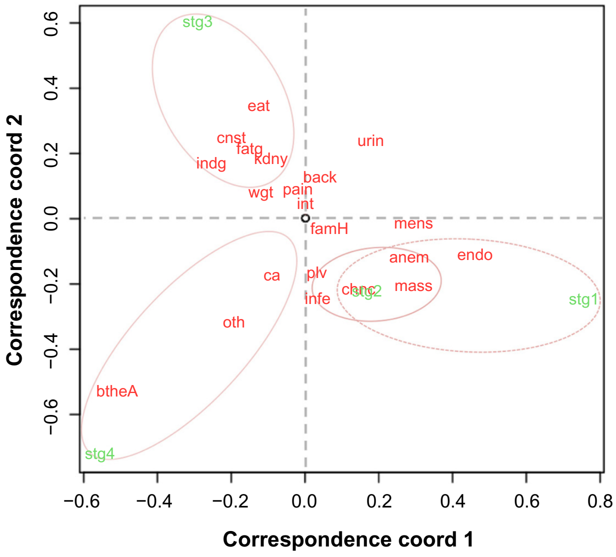


Figure 5. Correspondence analysis of stages and symptoms. Stage 1 is highly associated with a prior history of endometriosis. Note that the points close to the origin (0,0) have no particular strong association with anything else. See Table 2 for the definitions of the abbreviated symbols for prediagnosis symptoms.

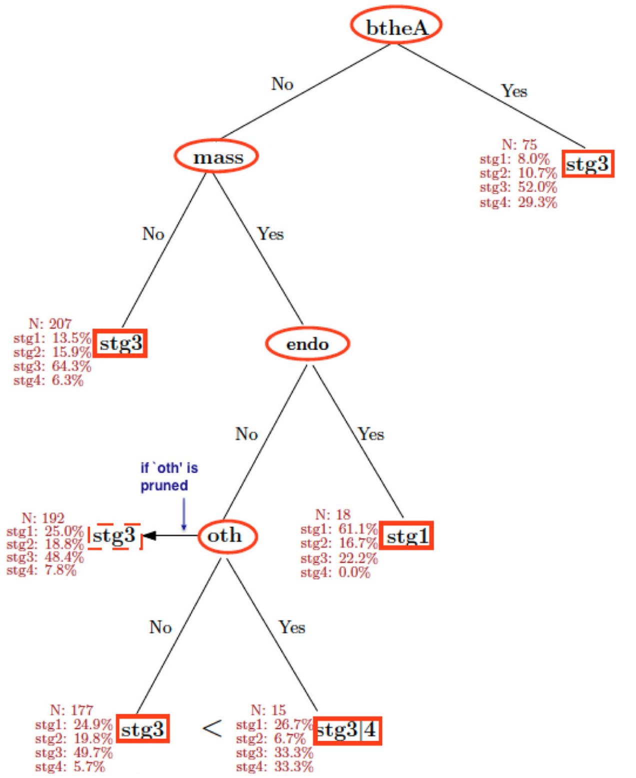


Figure 6. CART analysis of prediagnosis symptoms on diagnostic stages. The presence of endometriosis together with a mass is a significant predictor in separating Stage 1 diagnosis from other stages. See Table 2 for the definitions of the abbreviated symbols for prediagnosis symptoms. The visual representation of CART with its inverted tree structure is the same as that of a standard binary tree used in the "Instructions for Form 1040 (12/2013)–Internal Revenue Service": to decide if a taxpayer is qualified for a credit based on his/her response of Yes or No to a hierarchy of questions. In this figure, each split of the tree (eg, btheA) denotes a question (eg, Is the symptom of *breathing* difficulty and/or *fluid* [*ascites*] in chest or abdomen present?). The answers Yes and No to "btheA" would branch to two "daughter nodes": "Yes" to the terminal node "Stg 3," and "No" to another split "mass." This daughter, terminal node "Stg 3" on the right upper corner shows that Stage 3 is the best predicted diagnostic stage (with a 52% chance vs. a 25% chance if all four stages were equally likely under this node), while its annotated four frequencies indicate that Stage 4 is the second probable diagnostic outcome (with a 29% chance). At another daughter node, the split "mass," an answer "Yes" would branch further to the next-generation daughter node "endometriosis," to which an answer "Yes" would lead to the most probable diagnostic outcome, Stage 1. See Results for the explanation of the node "oth," the "other" symptom.

would be simpler: it would just lead to the most common diagnosis, Stage 3, without the node 'other' and its two subsequent branches. Without the pruning, we see that the resulting diagnosis (with an equal high chance of Stage 3 or 4) from a symptom *other* is worse than that (with a high chance of Stage 3) from no *other* symptom, after having a *mass* but not a presence of endometriosis. We used RF to rank the symptoms by their importance for classifying each stage and for the overall classification/discriminating accuracy (Fig. 7), by aggregating 5,000 trees built from fairly balanced subsamples of sizes (50, 50, 62, 50) for four diagnostic stages with out-of-bag (OOB) error estimates. We gave a slightly higher weight to Stage 3 as it is the most commonly diagnosed stage and this particular combination of subsample sizes produced an overall best classification rate among all different combination of subsample sizes we tried. The findings in Figure 7 are consistent to our findings in Figures 5 and 6. In Figure 7, the most important variables for classifying each of Stages 1–4 are endometriosis, eating, mass, and breathing + ascites, respectively, while the top-ranked variables for the overall accuracy of classification are endometriosis, mass, breathing + ascites, and menstrual problems.

Discussion and Limitations

The US Preventive Services Task Force (USPSTF) has recently recommended against screening for OvCa in women.[32] Although the report acknowledged that the mortality rate for OvCa is very high, their recommendation was based upon the low incidence rate among asymptomatic women. Moreover, the current standard diagnostic techniques were costly and invasive, such as transvaginal ultrasonography. The task force concluded that routine screening would produce a high number of false-positive results leading to unnecessary procedures, including interventional surgery. However, the USPSTF did not consider clinical studies on the use of symptoms together to guide detection of OvCa, because these were outside the scope of the commissioned systematic evidence review used to inform the recommendation statement.

Of the OvCa patients, 63% are diagnosed at Stage 3 or above. Although the incidence is low, the cost of care for OvCa patients is among the highest in the United States.

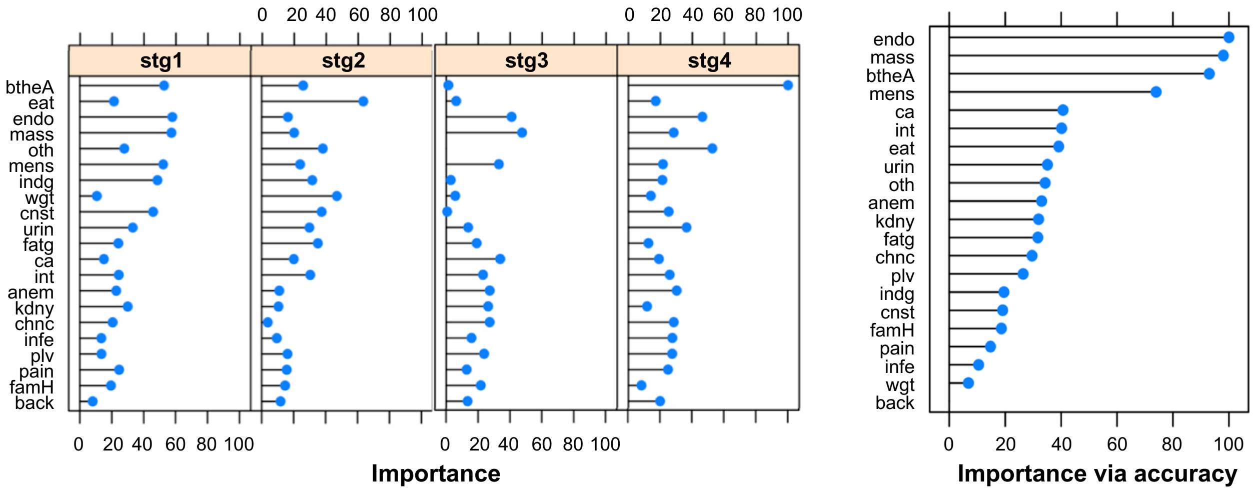


Figure 7. Variable of importance by randomForest for classification of each stage and for overall accuracy. Leftmost four panels list the variables that are important for classifying each stage, while the right panel lists the variables important for an overall accuracy of classification of all stages. For both randomForest and Caret packages, endometriosis is the most important variable among Stage 1 participants, and the overall accuracy also specifies endometriosis, mass, breathing + ascites, and menstruation problem as the most important ones among all. The number of trees evaluated was from 500 to 10,000, and 5,000 trees was the most stable in terms of OOB estimate of error rate. Caret package has the randomForest imbedded in.

It is therefore of value to explore potential relationships of nonspecific symptoms to enhance early diagnosis of OvCa. This approach has the potential to impact both patients and physicians. OvCa is a collective term for several distinct diseases, which can result in conflicting information being provided under a single all-encompassing terminology. As shown in all three statistical analyses in Figures 4–7, endometriosis has a stronger association with Stage 1 diagnosis. Indeed, numerous articles have shown that endometriosis has an increased risk to be associated with specific types of OvCa, see, for example, Van Gorp et al.[33] This sheds light for future development in possible prevention of certain subtypes of OvCa, or generation of a hypothesis about a possible diagnostic indicator of OvCa.

Although we cannot guarantee that our survey data are 100% free of recall bias, our survey design has implemented the best strategies as recommended by Hassan[27] and ECDPC[26] to prevent or minimize recall bias for a patient-reported study, such as, "improving timeliness of information gathering with more new diagnosis, framing questions to aid accurate recall with instructions, blinding the study subjects to the specific factors being studied (using free text)." In addition, we have cross-examined the text responses. Since our survey population has very few young (younger than 19 years old) and old (older than 80 years) participants, our research finding may not be applicable to these two extreme groups.

Conclusion

This work is a testimony of the power of team science by researchers from multiple scientific areas together with stakeholders. This study showcased our comprehensive survey of OvCa survivorship and methodology in design and implementation of a sensitive, responsive, and effective Web-based survey. It marked the debut of the OVA-CRADLE™ developed for Web-based access and online query and exploration of our comprehensive data. Our data and OVA-CRADLE™ facilitated subset of data, test, and generation of hypotheses. A study of prediagnosis symptoms of OvCa was accomplished using a combination text mining and statistics.

The top five symptoms across all stages were found to be a mass in the abdomen, eating and indigestion problems, menstrual irregularity, and pelvic pain (Fig. 4). For early stages, a mass in the abdomen and chance diagnosis such as an MRI after a car accident were the most likely diagnostic indicators of OvCa, while breathing problem or fluid (ascites), or a mass with other symptoms were strongly associated with late-stage diagnosis. Further investigation through CA of the prediagnosis symptoms with their relationship to diagnosis stage suggested that endometriosis was strongly associated with Stage 1 diagnosis (Fig. 5). CART analysis and randomForest also demonstrated an association between endometriosis and a mass in the abdomen for early-stage diagnosis (Figs. 6 and 7).

The application of our valuable data and database can go beyond the studies presented above. They can be used to test and generate additional hypotheses.

Acknowledgments

Thanks to Dr Vivian von Gruenigen and many OvCa patient partners for their feedback or testing of our initial survey. Some of our OvCa partners are no longer with us. Without the overwhelming support and participation of all these remarkable women, this research would not be possible. We also thank the wonderful reviewers whose comments have helped in improving the manuscript significantly.

Author Contributions

Conceived and designed the initial survey: JS, KMB, GQZ. Reviewed and participated in the design and implementation of the survey: JS, JT. Database, Webpage, and Unix scripting writer: JT. Developed the 20 key terms: JS, YHS, JT. Applied tm() in R: YHS. Wrote code and conducted initial statistical analysis: JS. Analyzed the more complete data: JS, RC. Wrote the first draft of the manuscript: JS, KMB. Contributed to the writing of the manuscript: JS, KMB, GQZ, JT, RC. Agree with manuscript results and conclusions: JS, KMB, GQZ, JT, RC, YHS, LC. Implemented OVA-CRADLE^{TM}: GQZ, LC. Jointly developed the structure and arguments for the paper: JS, KMB, GQZ. Made critical revisions and approved final version: JS, KMB, GQZ, JT, RC. All authors reviewed and approved of the final manuscript.

REFERENCES

1. NCI. 2014. Available at: http://seer.cancer.gov/statfacts/html/ovary.html. Accessed May 25, 2014.
2. American Cancer Society. *Cancer Facts and Figures* 2014. Atlanta: American Cancer Society; 2014.
3. American Cancer Society. 2014. Available from: http://www.cancer.org/cancer/cancerbasics/lifetime-probability-of-developing-or-dying-from-cancer. Accessed May 25, 2014.
4. NCI. 2014. Available at: http://seer.cancer.gov/csr/1975_2011/browse_csr.php?sectionSEL=21&pageSEL=sect_21_table.15.html#table1. Accessed October 15, 2014.
5. Carter RR, DiFeo A, Bogie K, Zhang GQ, Sun J. Crowdsourcing awareness: exploration of the ovarian cancer knowledge gap through Amazon mechanical turk. *PLoS One*. 2014;9(1):e85508.
6. Grady D. Widespread Flaws Found in Ovarian Cancer Treatment. *The New York Times*. Health Section, March 11, 2013. Available at: http://www.nytimes.com/2013/03/12/health/ovarian-cancer-study-finds-widespread-flaws-in-treatment.html. Accessed March 12, 2013.
7. Blanch-Hartigan D, Forsythe LP, Alfano CM, et al. Provision and discussion of survivorship care plans among cancer survivors: results of a nationally representative survey of oncologists and primary care physicians. *J Clin Oncol*. 2014;32(15):1578–85.
8. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010–2020. *J Natl Cancer Inst*. 2011;103(2):117–18.
9. Goff BA, Mandel L, Muntz HG, Melancon CH. Ovarian carcinoma diagnosis. *Cancer*. 2000;89(10):2068–75.
10. Available at: http://stat.case.edu/ovac. Accessed June 1, 2014.
11. Patient Reported Outcomes Measurement Information System (PROMIS). 2014. Available at: http://www.nihpromis.org. Accessed May 25, 2014.
12. Functional Assessment of Chronic Illness Therapy (FACIT). 2014. Available at: http://facit.org. Accessed May 25, 2014.
13. Available at: http://www.usatoday.com/news/nation/environment/2009-06-23-epa-study_n.htm. Accessed May 25, 2014.
14. EPA. 2014. *2002 National-Scale Air Toxics Assessment*. Available at: http://www.epa.gov/ttn/atw/nata2002/tables.html. Accessed May 25, 2014.
15. Zhang GQ, Mueller R, Szymanski J, T Adam, Wilson D, Flask C, Muzic R. Multi-modality, multi-resource, information integration environment. US Patent, 8,856,169, 2014.
16. NIH News. *Three New Informatics Pilot Projects to Aid Clinical and Translational Scientists Nationwide*. 2009. Available at: http://www.nih.gov/news/health/jan2009/ncrr-26.htm. Accessed May 25, 2014.
17. Zhang GQ, Siegler T, Saxman P, et al. VISAGE: a query interface for clinical research. *AMIA Jt Summits Transl Sci Proc*. 2010;2010:76–80.
18. Center for SUDEP Research. Available at: http://sudepresearch.org. Accessed January 20, 2015.
19. Zhang GQ, Sahoo SS, Lhatoo SD. From classification to epilepsy ontology and informatics. *Epilepsia*. 2012;53(suppl 2):28–32.
20. Sahoo SS, Lhatoo SD, Gupta DK, et al. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. *J Am Med Inform Assoc*. 2014;21(1):82–9.
21. Zhang GQ, Cui L, Teagno J, Kaebler D, Koroukian S, Xu R. Merging ontology navigation with query construction for web-based Medicare data exploration. *AMIA Jt Summits Transl Sci Proc*. 2013;2013:285–9.
22. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM*. 1975;18:613–20.
23. Heiler M, Schnörr C. Controlling sparseness in non-negative tensor factorization. In: Leonardis A, Bischof A, Pinz A, eds. *Computer Vision. ECCV* 2006*, Vol* 3951 *of Lecture Notes in Computer Science*. Berlin, HD: Springer; 2006:56–67.
24. Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction. *J Mach Learn Res*. 2003;3:1083–1106.
25. Feinerer I, Hornik K, Meyer D. Text mining infrastructure in {R}. *J Stat Softw*. 2008;25(5):54.
26. ECDPC. *Field Epidemiology Manual by European Center for Disease Prevention and Control (current as of Dec 10, 2014)*. 2014. Available at: https://wiki.ecdc.europa.eu/fem/w/fem/preventing-bias.aspx.
27. Hassan E. Recall bias can be a threat to retrospective and prospective research designs. *Internet J Epidemiol*. 2005;3:2.
28. Howden LM, Meyer JA. *Age and Sex Composition: 2010. 2010 Census Briefs, US Department of Commerce, Economics and Statistics Administration*: US Census Bureau; 2010. Available at: http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf [Table 2].
29. Blair AR, Casas CM. Gynecologic cancers. *Prim Care*. 2009;36(1):115–130.
30. Cancer Treatment Centers of America. *Ovarian Cancer Risk Factors*. 2014. Available at: http://www.cancercenter.com/ovarian-cancer/risk-factors.
31. Available at: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#features.
32. U.S. Preventive Services Task Force. *Screening for Ovarian Cancer*. 2012 Available at: http://www.uspreventiveservicestaskforce.org/uspstf/uspsovar.htm.
33. Van Gorp T, Amant F, Neven P, Vergote I, Moerman P. Endometriosis and the development of malignant tumours of the pelvis. A review of literature. *Best Pract Res Clin Obstet Gynaecol*. 2004;18(2):349–71.
34. Halpern MT, Ward EM, Pavluck AL, Schrag NM, Bian J, Chen AY. Association of insurance status and ethnicity with cancer stage at diagnosis for 12 cancer sites: a retrospective analysis. *Lancet Oncol*. 2008;9(3):222–31.
35. Yancik R. Ovarian cancer: age contrasts in incidence, histology, disease stage at diagnosis, and mortality. *Cancer*. 1993;71(S2):517–23.