CHAPTER 4:
PROSTATE CANCER DATABASE FEASIBILITY ASSESSMENT

Background
Prostate cancer is the most common form of non-skin cancer in men in the United States (American Cancer Society website, 2002) and is the second leading cause of cancer-related death among men in the United States (Jemal et al., 2002). The American Cancer Society (ACS) projects that in 2002, 189,000 new cases of prostate cancer will be diagnosed. Approximately 70 percent of all cases are diagnosed in men age 65 years and older. Prostate cancer is about twice as common in African-American men as in white men. The incidence of prostate cancer in the United States has increased over the last decade, owing to a combination of increased awareness among men and the use of serum PSA testing for early detection of prostate cancer.

As noted by the Prostate Cancer Outcomes Study (PCOS) group of the National Cancer Institute, patients with prostate cancer face uncertainty from many issues surrounding the management of the disease.

_It is not known, for instance, if the potential benefits of prostate cancer screening outweigh the risks, if surgery is better than radiation, or if treatment is better than no treatment in some cases. Decisions about treatments are also difficult to make … there are no randomized trials that compare the relative benefits of treating early-stage patients with radiation therapy, radical prostatectomy (surgical removal of the entire prostate gland along with nearby tissues), or watchful waiting. About 80 percent of men diagnosed with prostate cancer have early-stage disease. (National Cancer Institute PCOS website, 2002)_

The therapeutic options for prostate cancer include watchful waiting, radical prostatectomy (surgery to remove the prostate), brachytherapy (radioactive seed implantation), external beam radiation, orchietomy (removal of the testes) and androgen deprivation. Specific treatment—radiation therapy, radical prostatectomy, and hormonal therapies—can have serious effects on urinary, bowel, and sexual functions. Alternative treatments, such as diet modification, exist as well, although few alternative therapies have been evaluated. However, many patients with prostate cancer report using alternative treatments in concert with a traditional form of treatment.

Given the lack of clear scientific information to guide treatment choices for patients with prostate cancer, this condition was viewed by the study team and AHRQ
as an excellent choice for assessing whether a national longitudinal prostate cancer outcomes data base would be a valuable tool to assist with treatment decisions and for evaluating the feasibility of creating such a data base.

The feasibility assessment consisted of identifying a set of questions to which patients and providers want answers (Appendix A), examining existing data bases and longitudinal data collection efforts in the area of prostate cancer, engaging providers and patients in a discussion of the utility of the data base concept and issues related to the feasibility of collecting and organizing data that could be used by patients and their physicians, and identifying key issues that should be considered in the design of a data base.

Consultative Meeting with Prostate Cancer Patient and Provider Representatives

We began the feasibility assessment with a set of one-on-one phone conversations with patients and/or representatives of patient groups and providers to explore the need for a People Like Me data base for prostate cancer, to explore what information was desired and where information gaps exist, and to identify existing data collection efforts that are tracking prostate cancer patients longitudinally. From these discussions, we identified a set of physicians and prostate cancer patient representatives whom we then invited to a one-day consultation meeting in Rockville, MD.

On March 22, 2002, RAND and AHRQ convened a group of 13 representatives (see Appendix C for a list of participants)—prostate cancer patients, prostate cancer advocates, and researchers and physicians who had been directly involved in collecting and disseminating information on outcomes of care. The purpose of the meeting was to obtain the input of potential end users—patients and providers—on the idea of developing a People Like Me outcomes data base on prostate cancer treatments and to explore major data base design issues. During the meeting, we asked the group to do the following:

• Define the important questions to which patients and providers want answers, so as to understand what data elements need to be captured and how many patients should be recruited to participate in the data base (see Appendix A).
• Define how the data would be used—what pieces of information are needed by whom and in what form.
• Evaluate how willing patients and/or providers would be to allow their data to be entered into the data base, under what conditions, and how potential barriers to participation might be addressed.
• Assess whether existing data bases provide a potential platform on which to build a national outcomes data base (could they be modified or enhanced?) and opportunities for collaboration (lessons learned).
• Discuss problems and challenges (e.g., confidentiality concerns, bias if full population of patients is not represented).
• Explore what entity would be a respected and trusted sponsor of such a data base.

For a portion of the meeting, participants were divided into two groups (patient/advocates and providers), and a RAND researcher led the discussion in each group. We decided to separate the patient and provider groups because we wanted each stakeholder group to have an opportunity to freely discuss the topics from its individual perspective. Below we summarize the discussion from this meeting.

Is There a Need for a National Outcomes Data Base for Prostate Cancer?

Overall, both the patient and provider representatives felt that a national prostate cancer outcomes data base would be of significant benefit to both patients and providers. The patient representatives indicated that it would be very helpful to be able to access information on various treatment options and side effects that fit their specific characteristics (e.g., age, race/ethnicity, stage of cancer). The patient group reinforced their belief that current access to information on prostate cancer and treatments is piecemeal at best and the information that is available is sometimes contradictory. The patient group also observed that the type of provider seen usually dictates the type of treatment recommended (i.e., urologists generally feel that radical prostatectomy is the best treatment course, whereas radiation oncologists lean toward radiation). Although many men diagnosed with prostate cancer know there are different treatments available, they are not sure where to turn to learn about the difference in outcomes for the various treatments. Uniformly, the patient advocates stated that a national data base would provide an independent place that patients and their families could turn to for up-to-date information that would help them make an informed decision on which type of treatment to pursue.

Providers generally felt that the data base would provide them with a tool to fill in knowledge gaps related to prostate cancer treatment, such as how certain people respond to treatments and/or how the treatment affects functional status and resumption of normal activities. In addition, a national data base might allow for unique populations to be identified, such as those with genetic predisposition to prostate cancer. Both groups agreed that a national outcomes data base would provide
an important tool for shared decision-making between patients and providers, something that is currently lacking.

**Existing Prostate Cancer Data Base Efforts—Do They Serve as a Starting Point?**

A number of prostate cancer outcomes data bases have been created and are currently operational. During the meeting, representatives from three of these data base efforts—the Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE), the Tri-Service military Center for Prostate Disease Research (CPDR) registry, and a hospital-based system at Memorial Sloan Kettering—gave brief descriptions of the content and uses of their data bases. CaPSURE collects data on prostate cancer treatment outcomes at 32 sites across the country and has over 8,000 men enrolled. CPDR (based at Walter Reed Medical Center) is a data base of over 16,000 men in 11 military hospitals. The third is a small and growing data base established by clinicians and researchers at Memorial Sloan Kettering and was designed to produce statistical prediction models to help guide treatment decisions. Detailed summaries of these data bases and other data collection efforts follow later in this chapter.

Although these data bases and others, such as the Surveillance, Epidemiology and End Results (SEER) program and the PCOS, are important tools for prostate cancer research, each is primarily geared to surveillance or research, and not to direct access by a patient’s physician who is making decisions on their behalf. Currently, none of these data base efforts allow patients to directly access information or to tailor information from the data base to their specific characteristics to help guide decisions about treatment. So while much good information has been collected, little of it is being shared directly with patients who face the challenge of making treatment choices.

The discussion among the meeting participants also noted that each of the data bases has its own set of limitations, which prevents any single data base from serving as a platform or preexisting source of data to fully populate a People Like Me data base. All data bases would have to be modified. Examples of the limitations of existing data bases include

- missing subgroups of men (i.e., the poor, those with no health care, African-Americans, Asians, and those obtaining care from community practicing urologists)
- focusing on the initial course of treatment with no follow up information on subsequent courses of treatment or outcomes of care other than mortality (i.e., SEER)
• not capturing the full range of treatments by focusing data collection on urologists and not directly engaging radiation oncologists in the data collection
• focusing data collection on large centers of care and potentially missing a portion of the population that is different—both in patient characteristics and the type of treatment received in smaller communities
• not making data on outcomes directly available to patients or physicians for use in treatment decisions.

The Prostate Cancer Outcomes Study comes closest to capturing information on a representative sample of the population of prostate cancer patients; however, the data from this study could be better used to draw out important outcome information that could be shared with clinicians and patients.

Participants at the meeting agreed that the national outcomes data base could be modeled after these current data bases but should be expanded and would need to be made more patient friendly—specifically, that information from the data base would be made available to consumers and packaged for their use. Participants also noted that more information on pre- and posttreatment functional status is needed across all of the data collection efforts to determine the net effect of treatment, and that it would help if these assessments were standardized across the data bases to permit comparisons. Because there are variations across the different data collection efforts in the data elements collected, data definitions, and outcomes assessed, pooling of data from existing data collection efforts for the purposes of meta-analyses is also problematic.

What Information Is Needed?

During the meeting, we asked the two groups to comment on a draft set of questions an outcomes data base should focus on (see Appendix A). Overall, the physician and patient representatives agreed that the list of questions was appropriate. In particular, patient representatives expressed a desire for information on each step of the diagnosis and treatment process—for instance, what a rising PSA means, what the chances are of returning to pretreatment functioning, how the various treatments affect quality-of-life, and what treatments are available for complications/side effects of the cancer treatment.

Both groups agreed that a national prostate cancer data base would benefit patients and providers in terms of selecting treatment options and managing side effects. Both groups felt that although there is a great deal of information available on prostate cancer, many patients and providers do not have access to the most current information on the outcomes of various treatment options. For example, information on
side effects and posttreatment functioning are based on follow-up studies from patients who were treated ten years ago, and these results may reflect very different practice styles and treatments than are currently used. Consequently, the quality-of-life outcomes of patients who were treated 10 years ago when there were fewer options available or when treatments were more invasive, are not necessarily applicable to patients undergoing treatment today.

Both groups agreed that data are not available on how prostate cancer treatment affects general aspects of functioning and quality-of-life, such as being able to complete normal daily living activities and returning to work (Volk et al., 1999). Most data currently available focus on urinary/bowel dysfunction and sexual dysfunction. Patients who had undergone treatment stated that they did not realize that it could take up to a year to recover to pretreatment functioning and they were under the impression that it would take a matter of weeks. In addition, in terms of treatment success, patients want information on gauging how successful their treatment is in terms of functioning (i.e., what constitutes successful treatment). For example, if a patient is suffering from postsurgery incontinence, how many pads per day is normal for a patient who has been “successfully” treated? Finally, patients stated a desire to be able to learn who the top physicians are who perform prostate cancer procedures. Patients expressed a desire to make an educated decision on whom to see for the various treatment procedures (surgery, hormonal therapy, radiation). They wanted to know which doctors and which hospitals had done the largest volume of procedures, and—optimally—what the outcomes of care were for patients treated by different physicians.

**Participation, Sponsorship, and Data Collection**

There was consensus between both groups that a prostate cancer data base would be useful for both patients and providers and that data should be collected with input from both patients and providers. The patient group did not feel it would be difficult to get men to agree to participate and that confidentiality would not be an issue. One patient advocate had done an informal survey of over 100 men from a local prostate cancer support group: All agreed that a national data base is needed and would be beneficial, and that they would participate without hesitation. When asked who should sponsor such a data base, the patient group agreed that it should be a nonpolitical, highly visible, trusted body, such as the American Cancer Society.

When asked how the information in a national data base should be collected and accessed, many agreed that a Web-based design would be good, but that it might miss groups of men, such as those who are not computer literate and those who do not have access to a computer. Patients thought that a multifaceted approach to gathering and
providing information, such as through the computer and through written materials, would be best.

The most important aspect of a national prostate cancer data base, as cited by both patients and providers, is that it would provide a tool by which both patients and providers could make an informed, shared decision on which treatment options are best for a particular patient. This decision would be based on the patient’s age, state of cancer, PSA, Gleason score, grade of biopsy, and comorbidities. Patients would be able to evaluate differences in functional outcomes of various treatments by patient race, income, and geographic area.

Summary of Existing Outcomes Data Bases

CaPSURE.net

Overview

The Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE) is a research project managed by the Urology Outcomes Research Group of the University of California at San Francisco (UCSF) Department of Urology. It is sponsored financially by TAP Pharmaceuticals. The project has built a data base that collects longitudinal data on prostate cancer patients located across the United States. The data base collects information on treatment outcomes and the impact of prostate cancer on a patient’s quality-of-life. Researchers and policy makers at UCSF use the data base to evaluate outcomes of prostate cancer. Information is fed back electronically to physician participants in the study to track how their patients and practice patterns compare to those of other urologists. In some cases, although this is not the norm, physicians report using the information with their patients to help make treatment decisions.

Patients were first enrolled in the data base in 1995, and all data collected between 1995 and 1999 were manually collected and entered. In 1999, data entry and access to the data base became Web-based. This allowed for more patients to participate without substantially increasing costs. It also gave physicians the ability to see their patients’ data and compare it to national data in real time.

A Gleason Score is a grading system used by pathologists to describe the appearance of the prostate cancer tissue. It provides the physician with information to make estimates regarding the length of survival following a diagnosis of prostate cancer, and is generally combined with a patient’s PSA level and clinical stage to make a survival estimate. The Gleason grading system is based exclusively on the architectural pattern of the glands of the prostate tumor. It evaluates how effectively the cells of any particular cancer are able to structure themselves into glands resembling those of the normal prostate (a tumor whose structure is nearly normal—well differentiated—will probably have a biological behavior relatively close to normal). The grading system runs from grade 1 (very well differentiated) to grade 5 (very poorly differentiated). Grades 1 and 2 closely resemble normal prostate cancer cells whereas grade 5 usually predicts a poor prognosis. The Gleason Score is the result of summing two Gleason grades from different specimens. Thus the lowest possible Gleason Score is 2 and the highest possible is 10.
Currently, 8,100 patients are enrolled in the data base, and about 1,500 new patients enroll each year, a number that is constrained by the budget. Approximately 5 percent of the patients drop out each year. Physicians at 32 practice sites participate in the study by entering their patients’ clinical data into the data base. Nurses and other clinic staff also assist in the task of entering clinical data. The majority of physicians participating in the project are urologists, so the data base is heavily weighted toward patients who undergo surgery. The practice sites are located throughout the country; however, the patients and physicians within the CaPSURE sites do not represent the population of patients receiving care for prostate cancer or the population of physicians and facilities that provide care to prostate cancer patients. The vast majority of data contained in the data base come from urologists in selected practice sites, and the inclusion of radiation oncology information and patients who undergo this treatment is limited.

**Data Base Components**

The data base contains two types of information: (1) clinical data on patients and (2) patient-reported data on quality-of-life. Physicians and staff at the participating clinics enter clinical data for individual patients onto a password-protected secure web-server hosted by a third party (NetOutcomes). The clinical data collected include demographic information, method of cancer diagnosis, clinical pathological staging, PSA values, Gleason scores, treatment choices, and various lab results. The website for the data base has several features that enable the physician to do analyses on the patients’ data and compare it to national data. The analyses are displayed in graphical formats and do permit subgroup analyses (such as demographics, and diagnosis and staging).

CaPSURE also contains patient-reported data on quality-of-life issues from semiannual mailed surveys. The quality-of-life issues that are addressed in the surveys include treatment side effects, such as urinary, bowel, and sexual function, satisfaction with treatment, fear of recurrence, physical functioning, bodily pain, general health, fatigue, social functioning, and mental health. Researchers and site management staff at UCSF manage the quality-of-life survey process. The data from returned surveys is de-identified before it is transmitted to the third-party web host and made available to physicians. Therefore, physicians see only aggregate quality-of-life data provided by all patients in the clinic, rather than an individual patient’s responses. As with the clinical data, the data base offers several analysis functions so that physicians can compare quality-of-life information for all patients in their clinic with national data.
By viewing the information within the data base, physicians can show patients information about side effects associated with the different types of treatment. They can also look at the characteristics of patients who receive a particular type of treatment, but because the data base does not include radiation oncologists as a primary source of data, their patients are underrepresented in the data base. The semiannual survey asks patients to report on whether they used some type of complementary or alternative medicine to treat their cancer, though it does not capture detailed information to examine what alternative therapies are being used and with what effect. About 30 percent of the patients in the data base report that they use some type of complementary or alternative medicine.

**Physician Participation**

The physician sites that participate in CaPSURE are a convenience sample of urology practices. In the first two years of the project, the study team recruited sites known to have large urology practices and also included sites from the members of the project’s advisory group (whose urologists were interested in the study). Once the program became operational and known, the recruitment effort largely shifted to a strategy where urology practice sites self-identified themselves as being interested in joining the project. The CaPSURE team then selected the sites on the basis of their ability to recruit patients and their geographic or site diversity. There has been growing interest among urologists in participation in the CaPSURE data base. In the past three years, the major selection criteria have focused on filling gaps in site, ethnic, and geographic variation to increase participants from the Rocky Mountain and Midwest regions.

When a clinic is selected to participate in CaPSURE, all physicians who treat prostate cancer patients at the clinic are invited to participate in the data base project, though not all of them do. Physicians are paid $150.00 for each new patient they recruit, and they receive $100 annually for each continuing patient. This payment does not fully cover the costs to the physician for participating, indicating that the participating physicians have a strong interest in the goals of the project.

Physicians use the information in the CaPSURE data base in a variety of ways. They can obtain aggregated summaries on patients in their clinic or aggregated summaries on all current and former participating sites. According to the UCSF staff, physician have used the information in their discussions with patients, although on a limited basis.
**Patient Characteristics and Recruitment**

Physicians and clinic staff at each practice site recruit patients who fit certain eligibility criteria. To be eligible to participate, patients must have a new, biopsy-confirmed prostate cancer diagnosis. Participating patients must provide written consent before being enrolled in the data base. If patients decline to participate, only limited demographic information is entered into the data base for research purposes (to compare participants to nonparticipants). Patients have the choice of participating in the project at different levels. They can consent to having only their clinical data collected, or they can choose to only complete the quality-of-life surveys, or they can consent to both. After treatment is completed, patients still receive the quality-of-life surveys every six months, regardless of whether or not they are still receiving medical care from the participating clinic. The average length of enrollment has been 36 months, although the length of time continues to increase.

There is a strong effort to include a diverse set of patients in the data base, though this has been a continuing challenge to researchers. Currently, 10 - 12 percent of the patients in the data base are African-American. The data base contains few Latino and Asian men. The patient survey is available in English and Spanish; no other patient group in the data base is large enough to justify the cost or effort to translate the survey into any other language. The average age of patients in the data base is 60 years old. Patients receive no payment for participating, but they do benefit by receiving information about prostate cancer research through a newsletter and by accessing publications online.

**Online Format**

The data base is accessible through a website that was developed by NetOutcomes. NetOutcomes is responsible for maintaining the website and data base and responding to all technical assistance inquiries from participating sites. Researchers at UCSF receive customized data sets from NetOutcomes on a quarterly basis.

The CaPSURE.net website has four major functions: (1) data entry; (2) clinic graphs; (3) research; and (4) site management. In the data entry area, individual patients can be found by specifying any of the following: doctor’s name, the patient’s first or last name, patient ID, patient medical record, patient date of birth. New patients can also be added. Once a patient is found, summaries of the following can be displayed: status, clinic visit, labs and imaging, prostate biopsy pathology, surgery and treatments, medications, erectile dysfunction, clinical trial, survey pathology, hospitalization summary.
The “clinic graphs” area of the website displays aggregate outcomes and demographics under the following headings: demographics, diagnosis and staging, treatment, recurrence and survival, and quality-of-life. These subareas are called “grouped patients analyses.” Most display outcomes and demographics as bar charts. The exception is “Quality-of-Life,” where percentages are plotted over time. Time is measured as months since treatment. For all grouped patient analyses, it is possible to graph either data for the individual clinic, for the aggregate data set, or for both. It is also possible to show a subset of the clinic data to display only outcome for one doctor. Most of the reports are “fixed” in their basic design but allow the end-user some flexibility to specify and create comparisons of interest.

Each grouped patient analysis consists of individual patient analyses. Treatment analyses include these individual analyses: initial treatment, prostatectomy method, hormonal therapy, luteinizing hormone–releasing hormone (LHRH) agonist use, anti-androgen use, and irradiation method. Quality-of-life consists of these individual analyses: urinary function, bowel function, sexual function, satisfaction with treatment, satisfaction with care, fear of recurrence, physical function, bodily pain, general health, vitality, fatigue, social function, and mental health.

Further subsetting of the clinic data is possible depending on the individual patient analysis chosen. For example, if one chooses “Prostatectomy Method” under the treatment analysis group, the data can be analyzed by insurance status, age at diagnosis, PSA at diagnosis, Gleason score, or tumor node metastasis stage. It is possible to display outcomes by conditioning on individual categories for any combination of these variables.

**Operational Aspects and Costs**

Currently, TAP Pharmaceuticals provides all funding for CaPSURE and owns the data contained in the data base. It does not own the data base software or data base development. The data base management is outsourced to NetOutcomes. NetOutcomes developed the CaPSURE.net interface and worked with the project’s researchers to convert the data base to the online format. Researchers at UCSF receive data dumps from NetOutcomes once every three months for the purposes of conducting research. TAP Pharmaceuticals is not involved in the research efforts.

Cleaning and pre-processing the data address problems arising from changes to specifications over time (common to longitudinal studies), shortcomings in the data base design (e.g. duplicate observations, missed opportunities for input validation), and different variable names used by UCSF programmers from those used in the data base.
Because they do not own the data base design, researchers at UCSF and TAP Pharmaceuticals have limited influence on design changes.

Costs to the project include labor for the UCSF staff, payments to NetOutcomes for hosting the data base and website on its server, and payments to the participating sites and physicians. We were unable to get an estimate of NetOutcome’s costs for its role in the project. A low estimate of the annual costs to UCSF is $3.5 million. The project budget provided by TAP Pharmaceuticals strictly determines how many patients can be enrolled in the data base. Since the sites enroll new patients each year, the costs for the sites are always increasing. Determining how many patients to enroll based on the budget can be difficult because the dropout rate throughout the year is unknown. UCSF is currently holding the enrollment of new patients to 1,500 per year.

**Staffing**

Coordinating and maintaining the CaPSURE data base require a multidisciplinary staff at several organizations. Staffing for the project includes employees of TAP Pharmaceuticals (the sponsor), the UCSF Urology Outcomes Research Group, and NetOutcomes.

The CaPSURE staff at UCSF comprises two teams, with 8-10 people on each team. One team includes the researchers who set research priorities for the project and determine what needs to be included in the CaPSURE data base. This team includes several SAS programmers, statistical consultants, three urologists who provide clinical expertise about prostate cancer, and researchers who conduct a variety of studies. One of the SAS programmers is an experienced data base programmer who supports the data transition from the data base at NetOutcomes to the needs of several SAS programmers at UCSF. His expertise is critical, in particular given that the NetOutcomes software is proprietary. Without this programmer at UCSF, data quality would suffer greatly, although this might not become obvious to the public, because data quality is not easy to measure.

The other UCSF team is a site management team and comprises 8 – 10 staff members. This team is involved in all aspects of managing the participating sites and spends a large part of their time on work related to the semiannual mailed patient surveys. Approximately four months out of the year is spent working on the surveys. The team is responsible for everything involved in the production and mailing of the survey, with the exception of printing, which is outsourced. The team also manages the survey data collection process. They track the returned surveys and follow up with patients if there are any questionable responses, including problems with illegible handwriting. When a survey is returned, the team scans the surveys into a computer to
create an electronic form of the survey, de-identifies it, and transmits the data to NetOutcomes. NetOutcomes then enters the survey data into the database and makes them available to physicians.

Converting to Web-based data collection for the patient survey would significantly reduce project costs and eliminate problems caused by illegible handwriting. However, the patient population at this time lacks sufficient access to the Internet and/or interest in completing the survey online. As access to the Internet increases and the patient population becomes more comfortable using the Web, providing the survey online will be revisited. Another factor that will need to be considered in evaluating the merits of shifting to online administration of the survey is the access speed of the Internet connections of patients being asked to complete the survey; patients with slower phone modem access may be unwilling to complete the survey online depending on its design.

NetOutcomes also has several staff members devoted to working on the CaPSURE database, although the exact number is unclear. These employees also have a diverse set of skills, including database management, website design, and usability testing. NetOutcomes is also responsible for responding to all technical assistance questions from the participating sites.

Challenges and Future Directions

The CaPSURE project, along with most prostate cancer outcomes databases, faces several challenges to being able to give highly personalized information about the range of prostate cancer treatments and outcomes to all patients. First, the database is not representative of all the providers of treatment; in particular, radiation oncologists are a minority in the database because the clinical sites targeted are urology practices. CaPSURE also does not have much penetration in the Rocky Mountain states, and represents a narrow swath of clinic sites. Additionally, the patient population is fairly homogeneous, both in terms of treatments and limited enrollment by minority populations. Each of these factors reduces the generalizability of the information. Staff at UCSF seem interested in making the database accessible to patients in the future but note that much work is required to determine what information should be provided and the format for doing so.

SEER Database

Overview

The Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute collects a wide range of information on all types of cancers diagnosed in
the United States. Prostate cancer is one of the many cancers included in the SEER data base. The data in the SEER data base come from population-based cancer registries that represent approximately 26 percent of patients with cancer in the United States. Though national in scope, the data base is not truly representative of all cancer patients in the United States.

The registry seeks to collect information on all cancer diagnoses that occur among residents within a geographical area. The data comes from a variety of sources including hospitals, physicians, and laboratories. Any health care provider or facility that serves patients who live in a registry area is mandated to provide data on residents’ cancer diagnoses. The majority of data comes from hospitals and is collected from claims data and medical record reviews. Because the SEER data base population is based on the residents in the SEER registry area, data may also be collected from hospitals and providers in neighboring states.

The number of registries has fluctuated since the SEER program began in 1973. Registries currently operate in Kentucky, Greater California, Louisiana, New Jersey, Georgia (rural areas plus Atlanta), Connecticut, Iowa, New Mexico, Utah, Hawaii, American Indians in Arizona, Alaskan natives in Alaska, and the Seattle-Puget Sound area. Recently, SEER has expanded its data collection to cover additional regions in an effort to address gaps in populations reflected in the data base.

The SEER registry tracks individual cancer diagnoses rather than individual patients. In other words, a patient is added to the registry when he or she gets an initial primary diagnosis of cancer in one area of the body. If the same person receives a primary diagnosis of cancer in another area of the body at a later time, the patient enters the data base again as a second case. Therefore, the registry may contain more than one cancer case for a single individual. It is not known how many actual people the registry represents.

The SEER registry conducts annual follow up for vital status only. Because the registry provides survival statistics, follow up with patients is continued until their death. The SEER program primarily gets data on date and cause of death from state death certificates. It also sometimes retrieves this information from Medicare sources through the Social Security administration. The validity of data from state death certificates and cause of death determination has been discussed and questioned in the scientific literature.

The SEER data base makes data on cancer incidence, mortality, and survival rates available in a public-use data file. There currently are data on three million in situ and invasive cancer cases, and approximately 170,000 cases are added to the registry each
year. The data available to the public are usually two years old and are de-identified to protect the identity of patients.

The annual budget for SEER is $25 million, which covers the registry staff as well as roughly 15 staff (not all working on SEER full-time) within the National Cancer Institute’s Statistics Branch who conduct statistical analyses of the data.

**Data Base Components, Uses, and Limitations**

The data base contains demographic data (age, gender, race), primary tumor site, morphology, age at diagnosis, stage of cancer at time of diagnosis, first course of treatment, and survival rates within each stage of cancer. It does not contain information on the details of treatment and whether subsequent treatment is provided to a patient, pretreatment clinical characteristics of the patient, and patient functioning before and after treatment. The sole outcome variable in SEER is survival.

The SEER data base is a valuable source of data for such clinical information as incidence and survival rates. SEER can also provide information on characteristics of people who choose a certain type of treatment. However, it cannot provide more qualitative information about side effects and impact on quality-of-life based on a patient’s treatment choice. To accomplish this, the National Cancer Institute and other organizations use SEER data to perform additional research. Some of these studies are called SEER Rapid Response Special Studies. One of the largest of these studies is the Prostate Cancer Outcomes Study, sponsored by the National Cancer Institute.

Another major project using the SEER data base has linked the data in the registry to Medicare claims data. The SEER-Medicare data provide valuable longitudinal information about cancer diagnosis, treatment, and follow up. They can be an especially good source of information on acute and chronic complications of cancer treatments. The data have been linked multiple times, first in 1991, and then in 1995 and 1999. Data linkage with Medicare data was accomplished for 93 percent of people in the SEER data base who are 65 years or older. The SEER-Medicare data linkage is a valuable resource for prostate cancer information. The generalizability of all the data in the SEER-Medicare data base is being studied and should be considered when reporting the data, particularly for cancers that have lower incidence in older age groups. Although extensive efforts are taken to de-identify SEER-Medicare data, a potential for identification remains, and thus, the data are not available to the public.

**Patient Characteristics and Recruitment**

Patients in the registry area who receive a primary cancer diagnosis are automatically entered into the SEER data base. They are not given the option of opting out. The registries are dispersed throughout the country and some registries were
specifically established with the goal of increasing minority representation in the data base. SEER data on race is collected from medical records and registration information. Hispanic ancestry is collected as a separate variable and is identified by applying an algorithm to detect Spanish surnames. SEER’s race data for patients who are not African-American or white are less valid, and none of the SEER race data have been externally validated.

**Staffing**

Each SEER registry employs 20 - 40 full-time employees, who are responsible for collecting the data for the registry. Staff at the National Cancer Institute, including administrators and researchers, also work for the SEER program.

**Prostate Cancer Outcomes Study**

**Overview**

The Prostate Cancer Outcomes Study provides important information on quality-of-life and treatment outcomes for prostate cancer patients. The National Cancer Institute (NCI) began conducting PCOS in 1994. The study collects data from patient surveys and medical records. Patients were recruited from SEER cancer registries in Connecticut, Utah, New Mexico, Atlanta, Los Angeles, and Seattle. Over 11,000 men were eligible for the study because they had been diagnosed with biopsy-confirmed primary invasive prostate cancer between October 1, 1994 and October 31, 1995. From this study population, 5,672 men were sampled and 3,533 completed the six-month survey, 12-month survey, or both. Eighty-three percent of the men who were eligible to complete the 24-month follow up survey did so. Ninety-nine percent of the 3,533 men had their medical records abstracted. The study sought to generate a representative sample in terms of age, race, geography, health care setting, and treatment choices. Certain age and racial/ethnic groups were over-sampled, though the entire study focused primarily on white non-Hispanic, Hispanic, and African-American men. This is a strength of this data base that was made possible by building off of the SEER data.

Quality-of-life surveys were sent to participating patients at six, 12, and 18 months after their diagnosis. The surveys were available in Spanish. Follow up is done annually to check on vital status and address changes, and to provide a patient newsletter. In fall 1999, a five-year follow up survey was conducted, and the potential exists to conduct a ten-year follow up survey as well. PCOS also includes data from patients’ medical records, such as patients’ PSA scores, clinical stage and grade, treatments, and specific hormone therapies given. Data on age at diagnosis, race/ethnicity, clinical stage and grade of tumor, and the initial treatment choice came
from SEER. Education and income data were mapped from Census data, matching by zip code blocks.

The PCOS survey focuses on the impact of prostate cancer treatment on urinary, sexual, and bowel function. The treatment options that it focuses on are surgery, radiation therapy, and hormone therapy. The study data can tell a variety of things about what kinds of treatments patients with certain characteristics are choosing, as well as the side effects that are experienced by patients who choose those treatments. Although the PCOS survey asks patients whether or not they engage in alternative therapies (e.g., dietary changes, herbal supplements, exercise), NCI is not evaluating the relationship between these supplemental therapies and treatment outcomes or quality-of-life.

The results that have emerged from analyses of the PCOS longitudinal tracking data underscore the potential value of developing People Like Me outcomes data bases. Some of the completed research has found important differences in the side effects experienced by patients who choose different treatments. Another study has found that PSA levels, Gleason scores, and age are the best predictors of the spread of the disease outside the prostate. Analysis of the PCOS data also show that age at diagnosis, pretreatment health status, comorbidity, and treatment choice are the most important factors contributing to patient outcomes—covariates that a People Like Me data base would want to use to display differential outcomes tailored to the demographic and clinical characteristics of patients. Each of the analyses derived from PCOS data lends credence to the need to capture information on outcomes by different clinical and demographic characteristics of the patient. However, the information gleaned from the PCOS analyses is not being communicated directly to patients and doctors in a way that promotes shared decision-making at the point of care; so while PCOS realizes part of the vision of a People Like Me data base, it has not taken the additional step of creating an information tool for use by patients and physicians.

**Staffing, Operational Aspects, and Costs**

The staff for the PCOS study include four people at the National Cancer Institute and an unknown number of people who work with the six SEER registries involved in the study. The PCOS staff at NCI performed the study under contract, rather than as a grant, which provided some cost-savings. To recruit and track approximately 3,500 patients, PCOS has spent an estimated $7 - 10 million since 1994.

**Challenges and Future Directions**

One way to make better use of the PCOS data is to increase efforts to disseminate the results directly to patients—although work is required to understand how to do this
effectively. Furthermore, PCOS has not focused much activity on dissemination and does not know how results of its studies have affected practice behavior.

A PCOS staff member noted that much could be done to repackage existing information and disseminate it more effectively to patients. This would involve determining which People Like Me questions have already been answered or could be answered by the existing data bases, grading this evidence, and determining where the information gaps are and how best to fill them.

It was also noted that it would be valuable to jointly agree on common definitions for the variables that are being collected across data bases. A factor hindering the pooling of data across these collection efforts is that elements are difficult to compare because of differing definitions used and outcomes tracked.

**Center for Prostate Disease Research Tri-Service Multicenter Prostate Patient Data Base Registry**

*Overview*

The Center for Prostate Disease Research (CPDR) Tri-Service Multicenter Registry collects information about prostate cancer from men in the TriCare military hospital system. The data base captures information from the major military hospital centers nationally. The data base collects clinical data, complications, and limited quality-of-life information from men with diagnosed cases of prostate cancer and those who have clinical indicators of prostate cancer risk (i.e., elevated PSA scores but no positive biopsy of cancer).

Unlike those data bases that have been developed in the nonmilitary sector, the nature of the military organization reinforces physician participation in the project and also contributes to the high level of participation and compliance by the patients. For men who meet the clinical criteria for inclusion, 99 percent agree to participate in the registry. Within the military system, active duty personnel must be seen by a physician twice a year. These biannual visits allow CPDR to capture ongoing information about patients who have prostate cancer or are at elevated risk. Retirees are not seen as often, but these individuals still maintain fairly regular contact with the military hospital system to receive care, which facilitates tracking patients longitudinally.

CPDR has adopted a participatory model in its design and operation. At each hospital site, CPDR enlists physicians to become collaborators on the research team, so that there is active commitment to and participation by all sites in the data base. Physicians within each of the participating hospital sites are encouraged to request data from the data base for their own research and to publish their findings. The TriCare
system has successfully integrated the project into the clinical setting and even uses it to help maintain contact with patients and involve them in their care. The data are used to generate short monthly reports at the hospital site level to use in tracking and managing patients.

The data base officially started in 1994 and collected data from only one military hospital at that time. Up to as many as 13 military hospitals have participated since the project’s inception. The current number of participating sites is 11 and includes 16,025 patients. The hospitals are all located in large cities; small military hospitals located in small cities or more rural settings are not participating in the data base. CPDR does not have information on the nonparticipating sites to assess whether there are key differences in the populations (e.g., age) at participating and nonparticipating hospitals. CPDR does receive requests from other hospital sites that are not participating to join the registry program; however, CPDR does not have funding to support expansion of the registry to other sites.

Until 1999, the only patients enrolled in the data base were men with diagnoses of prostate cancer. Beginning in 1999, the data base expanded to include patients with elevated PSA levels. CPDR is interested in tracking these patients who are considered at elevated risk because approximately 80 percent of them will have active prostate cancer at some later point in time. As of June 2002, approximately 16,025 patients were enrolled in the data base. Of these, approximately 12,000 have an active case of prostate cancer; the remaining 4,000 patients in the data base have an elevated PSA level with a benign biopsy.

**Data Base Components**

The CPDR data base is a relational Oracle data base. The data base is not directly accessible through the Web or by other means. At this time, there is no mechanism in place for physicians or anyone else to be able to get data in real time.

CPDR collects general and prostate cancer-specific data, such as age, race, PSA levels, tumor differentiation, Gleason scores, and staging information. The data base contains approximately 500 data fields in 48 tables that include registration, patient contact information, pretreatment diagnosis, cancer staging, treatment types, follow up and recurrence, quality-of-life issues, and many others. There are detailed data tables for each type of cancer treatment. Data are collected for standard clinical care, which includes hormonal therapy, radiation therapy, chemotherapy, and surgery. Each major treatment type has a subset of questions asked. The majority of the information in the data base is collected during the treatment stage.
Data for the prostate data base is obtained from three different sources. The primary source is the military’s Composite Health Care System (CHCS) data base, an electronic data base that collects comprehensive medical information (laboratory and pathology values, pharmacy, treatments, etc.) about all patients within the TriCare System. It contains demographic information about each patient, as well as some specific clinical information about all health conditions. Rather than asking the providers to enter the same data twice, CPDR imports data relevant for the prostate data base directly from the CHCS. The CHCS captures nearly all the data CPDR requires; however, it is supplemented by additional data collection from two other sources.

The second source of data is clinical notes in the patients’ charts. Each physician is required to maintain a hard-copy chart for a patient in addition to the CHCS electronic data record. Clinical data come only from the provider and not from the patient. The clinical data are collected by on-site data managers in each hospital who review the medical charts and manually enter information into the CPDR data base.

The third source of data is patient surveys that primarily focus on quality-of-life issues, although to date these have been limited in scope. CPDR views patient-reported data as subjective in nature and relies more on clinical assessment from the two previously discussed sources. Patients fill out a baseline survey at their first visit (either elevated PSA or actual positive biopsy of cancer, but pretreatment). The survey collects information on family history and clinical issues, such as PSA level, and pretreatment sexual, urinary, and bowel function, and other quality-of-life issues. Once treatment has commenced, patients are surveyed annually during their visits to the hospital. If a patient has not been seen for over a year, the data manager will mail the survey to him. The survey is a component of the project’s efforts to keep patients engaged in the longitudinal data collection by prompting patients to make an appointment with the urologist. If a patient says that he would like to make an appointment, the data manager will contact the patient directly to make the appointment and get him into the system for reassessment.

Survey responses are entered into the data base, and the hard copy of the survey is attached to the medical record. Because patients are seen so frequently and this population is very compliant in making and showing up for appointments, the mailed survey is not seen as a major source of information for the data base. Rather, the focus is on gathering information for the data base during the clinical visit because data managers typically get more complete information from patients during their medical appointments. The mailed survey is 16 pages long and considered burdensome for patients to complete; it is undergoing revision to reduce its length and to capture a
broader set of data on the patient’s functional status. The project staff meets annually to
discuss the project, including possible changes to the survey.

The data base does not routinely collect information on alternative treatments,
although physicians are free to note this in an open data field. Data on the use of
alternative therapies are not included in the reports to physicians and are not used for
research. The primary reason for excluding the capture of information on alternative
treatments from the registry is the difficulty of developing standardized definitions of
alternative treatments, which makes it difficult to compare alternative treatment
outcomes.

Patient Characteristics and Recruitment

All of the men in the data base have either an active case of prostate cancer or an
elevated PSA level. Patients can only be recruited for the project if they receive care
from one of the 11 participating military hospital facilities. Most of the patients in the
data base are in the military or retired from the military, although civilians who receive
care at participating facilities can also participate. All participants provide written
consent.

The average age of patients at the time of enrollment is 61. The average length of
enrollment is 5 - 6 years, and the reasons for dropping out of the data base are varied.
Many men drop out when their cancer is in remission and they are no longer receiving
any treatment. Surveys are still sent to data base participants regardless of their health
status. Deceased patients are dropped from active participation in the data base.
CPDR’s goal is to maintain contact with the men in the data base indefinitely.

The data base does not collect information on the income level of participants, but
because they receive military benefits, it is unlikely that a significant number of men in
the data base are below poverty level. The data base is likely to be somewhat more
representative of all income levels, because treatment for military servicemen is free, and
thus cost is not a barrier to patients seeking and receiving care. As a result, the CPDR
data base is likely to include more low-income patients, compared to private-sector data
bases. Another advantage of the TriCare System is that men within the system can see
any provider they would like. Approximately 17 - 18 percent of the patients in the data
base are African-American. Because of the United States military presence in the
Philippines, Filipino men are well represented in the data base, but other Asian groups
are not. The data base contains few Hispanics. There are some Native Americans in the
data base, but the numbers are too small for group-specific analyses.
**Physician Participation**

Because of the unique characteristics of the military setting, the cooperation of the physicians is expected and financial incentives to participate are prohibited. CPDR has taken steps to engage physicians in each of the hospital sites to be co-partners in the data collection and research activities of CPDR. In addition to establishing co-investigator roles for physicians in all sites, CPDR holds annual meetings with the clinicians to discuss data fields, to address problems/questions, and to assist with data analysis for research.

CPDR is actively committed to ensuring that the data base is relevant for physicians to use in improving the care that they deliver to patients with prostate cancer. Physicians use the data within the data base to help make treatment decisions and to compare their clinic’s performance with other participating facilities within the TriCare system. Each month, CPDR sends providers a hard-copy report that summarizes site-level data. The report provides information on their own patients’ progress, and physicians can use it to compare outcomes across different sites (but not across different physicians at their own site). This can help motivate physicians to provide better care. Physicians can submit a request to CPDR to get data for individual patients for use when making treatment decisions. The data can be put in a graphical format and used in discussions with the patient, although this does not occur with great frequency.

Physicians and/or staff automatically inform the on-site CPDR data managers about patient appointments with registry patients, and they give data managers access to the patient charts. Some providers participate actively in the project by using the data base to conduct their own research. However, most physicians never request additional data from the data base.

It should be noted, however, that CPDR currently is not set up to provide patients direct access to the information to compare treatment options and outcomes. While this has been discussed as a possible long-term goal, CPDR is concerned about a range of issues including legal problems in making patient data public (Institutional Review Board concerns about protecting patient privacy) and that statistics may be old and may not predict a patient’s real problem. It is likely that if patients had access to the data at some point in the future, CPDR would be required to re-consent existing patients in the data base, which would be extremely burdensome and difficult to do. At this point, information from the system is provided to patients in a more neutral fashion, by sending patients reminders or prompts to have their PSA checked and, in some cases, by physicians as they counsel their patients.
Staffing

Several groups of staff members play a role in managing and using the data base. Physicians participate in the data base project by helping to identify patients who should be recruited into the project. The majority of participating physicians are urologists and medical oncologists.

Data managers are CPDR employees who work on-site at the participating facilities. They represent the data base project in the clinical setting and play a major role. There are currently 18 data managers working at the 11 facilities (1 - 3 managers per site). Most data managers have a medical background. They are responsible for recruiting patients, securing patients' written consent, providing health education support to patients and their families, performing data entry, transmitting data to the central data base on a daily basis, assisting principal investigators with data requests, and serving as the local representative for the project to the Institutional Review Board (IRB). The health education that they provide to patients about prostate cancer comes from the CPDR data base and the National Cancer Institute. While consulting with patients, the data managers present the prostate cancer information in a neutral way to avoid influencing patients to choose one type of treatment course over another.

The on-site data manager contacts all men who receive care from one of the participating military hospitals and who meet the study criteria. The data manager meets with each of these patients, describes the data base project, and encourages the patient to participate. Patients must provide written consent in order for their data to be collected for the data base.

There are many benefits to having the data managers work on-site. They can recruit patients at the start of their diagnosis and treatment to capture important baseline information for assessing outcomes. They can respond quickly to identify and recruit eligible patients, and because of the regularity of patient encounters with the military health system, they have frequent access to patients to perform follow up assessments.

The other project staff members are CPDR employees who maintain the data base and provide data to the hospital sites. CPDR also provides technical assistance to the sites and meets with data managers and principal investigators on an annual basis to discuss the project and make any necessary changes to the protocol or survey. CPDR employees play an important role in ensuring the quality of the data within the data base. CPDR staff identify missing data and data with errors using a manual technique as well as with a software package integrated into the data base. They notify the data managers, who are then responsible for doing necessary follow up and correcting errors.
Military medical residents assist in this process. Every effort is made to have complete, accurate data, though sometimes it is not entirely possible.

**Operational Aspects and Costs**

The CPDR receives approximately $2 million per year from the federal government through the Nery M. Jackson Foundation for the Advancement of Military Medicine to operate the data base. CPDR receives an additional three to four grants per year, which total between $500,000 and $1 million. The budget pays for the CPDR staff salaries and the salaries of the on-site data managers. The budget also covers the costs of providing data to the study’s principal investigators, reports that are in addition to the standard, monthly data reports. Physicians are not paid for participating in the project, so the cost of operating this program is less than would be the case in the nonmilitary sector. All indirect costs (e.g., photocopying, travel) are also covered by the military, so do not contribute to the $2 million plus annual budget. The decision to expand the data base by enrolling more patients or adding another facility is based primarily on financial considerations. If additional money were available, more patients could be enrolled and CPDR could consider adding additional facilities.

**Challenges, Recommendations, and Future Directions**

One issue of interest is whether the CPDR data base could be merged into a national data base. Owing to the constraints of operating within the military structure, this would be extremely difficult if not impossible. CPDR staff anticipate that the military would not agree to share data with private-sector organizations. The main reason for this is the potential legal issues surrounding patient confidentiality and usage of the data, because the original patient consent for use of the information was limited to the objectives of CPDR. Combining the CPDR data with data from other data bases would require that each patient in the current data base give new written consent. There is concern that this consent process would lead to significant patient attrition.

A significant challenge to creating any outcomes data base, which was noted by the developers of CPDR, is the politics involved in meeting the needs and interests of the multiple participating organizations and sites. Success depends heavily on securing the cooperation and collaboration of all parties involved, and getting researchers, physicians, and hospitals to work together is time consuming, resource intensive, and requires commitment by all the players. Physicians are likely to need incentives to encourage their participation, such as data management and reporting tools that help them manage care for their patients and avoidance of redundant systems for capturing information.
All the facilities for the CPDR are located in large cities, and patients from small cities and rural areas are therefore not represented in the data base. It is conceivable that patients in these areas are different or that their care experiences are different. For example, military retirees might move to smaller towns and be less likely to receive care from one of the participating facilities. It is possible that these patients might therefore be older and have a different cancer profile, different treatment patterns, and different outcomes compared to those patients from the 11 major centers. It is unknown whether the care differs between participating and nonparticipating hospital centers and whether patients at nonparticipating centers experience worse outcomes. The data base may not be representative of all men nationally to the extent that military personnel are more physically fit than the general population of men with prostate cancer.

If the CPDR were to be replicated in the nonmilitary sector, CPDR estimates that the annual cost to develop and maintain such a data base would be three times as high or $6 million annually.

The CPDR data base provides one of the best examples of the organizational elements needed to develop and maintain a national data base on prostate cancer and outcomes, but this is partly a function of the unique characteristics of the military, which may not be replicable in the nonmilitary sector.

**Prostate Disease Patient Outcomes Research Team**

In 1989, the Agency for Health Care Policy and Research (AHCPR) funded the Patient Outcomes Research Teams (PORTs) to attempt to answer questions about the effectiveness and cost-effectiveness of available treatments for common clinical conditions (Freund et al., 1999). One of the PORTs funded by AHCPR addressed benign prostatic hyperplasia (BPH) (Wennberg et al., 1993). The project involved a group of researchers from Dartmouth, Massachusetts General Hospital, and the University of Massachusetts.

Each of the PORTs was obligated to conduct a formal literature review on what was known about the outcomes of treatment for the disease. The PORT projects were also to use readily available data—observational in nature—to advance the understanding of outcomes of care as applied to patients in everyday practice. The PORTs often used Medicare and Medicaid claims data, hospital cost report and discharge abstracts, and encounter data from health plans. The PORT approach stood in contrast to the traditional approach of conducting randomized clinical trials (RCTs) to answer questions regarding whether one treatment is more effective than another.
The PORT strategy for constructing an outcomes data base also contrasts with those previously described in this chapter. Rather than first collecting the data and then building the decision support tool, the prostate disease PORT built the decision aid first, followed by the prospective longitudinal data base to follow up patients according to their choice of treatment. This approach was designed to fill in gaps and to test specific hypotheses.

The team began by conducting a literature review and convening focus groups of patients to identify the outcomes of interest to patients. The research team then built a decision tree that included all of the outcomes relevant to patients, populating it with estimates (i.e., probabilities) of the chances that a given outcome will occur, according to patient subgroup if possible. The probability estimates were derived from an evidence-based review of the literature, analysis of existing data bases, and patient surveys to augment the data base with patient reports of outcomes (e.g., incidence of incontinence, impotence, disease recurrence). In some cases, the probability estimates for outcomes of interest were weak or missing.

Once the task of assigning probability estimates was completed, the team used this information to build a decision support tool to promote shared decision-making between physician and patient regarding choice of treatment. The decision aid took the best available information on outcomes estimates and presented prognostic information by subgroups of patients. The decision tool attempted to help patients make decisions that included outcome states for which, at the moment of decision, they would have had no direct experience. Where outcome information was weak or missing, the decision aid informed patients of the limits of scientific data.

The prostate disease PORT then built the longitudinal data base through systematic follow up of patients in clinical settings where the decision aids are routinely employed. The PORT approach differs from the RCT and the prostate cancer data bases described above in that the strategy for populating the probability estimates comes from people who have participated actively in the choice of treatment. In effect, the decision support tool is a mechanism for assigning patients to the treatments they prefer.

As noted in Chapter 1, Wennberg and colleagues (1993) propose an alternative strategy to RCTs for generating outcomes information, called the Preference Clinical Trial or PCT (i.e., patients choose among all treatments after being offered information about the risks and benefits of conventional and experimental treatments). The PCT approach (1) implements shared decision-making for all patients at a given center who are eligible for surgery, (2) helps patients come to a decision, (3) for those with preferences, enrolls them in the “preference arm” of the study, (4) for those who have no
preference, offer a randomized clinical trial. All patients are followed in the same manner. This approach, observe the researchers, would provide an opportunity to examine differences in outcomes according to preference or randomized designs, and thus would provide information on external validity as well as information relevant to the question of the role patient expectation plays in determining health care outcomes.

When the BPH PORT was operational in the 1990s, the costs for conducting research to build the synthesized data base and decision support tools were $1 million per year.

Memorial Sloan Kettering Longitudinal Patient Data Base

This data base is resident in the hospital and tracks patients longitudinally to generate data to run medical statistical prediction models (i.e., nomograms). Partin nomograms (Partin et al., 1997) are statistical models that use patient clinical characteristics (i.e., PSA level, clinical stage, and Gleason score) to predict pathological stage for men with localized prostate cancer. This is important because as many as 50 percent of men undergoing radical prostatectomy were found to have extraprostatic disease at the time of surgery, rather than organ-confined disease. Patients undergoing radical prostatectomy for localized prostate cancer who have organ-confined disease demonstrate markedly improved biochemical disease-free survival compared with men who demonstrate extraprostatic disease (Pound et al., 1997). The authors in this study note that clinicians can use these nomograms when counseling individual patients regarding the probability of their tumor being at a specific pathological stage, so that they can make more informed treatment decisions.

At Memorial Sloan Kettering Cancer Center (MSKCC), clinicians have been tracking clinical outcomes for their own patients who receive radical prostatectomy, external beam radiation, and brachytherapy. Their statistical models or nomograms use PSA value, Gleason sum, and clinical stage to help predict biochemical freedom from recurrence for radical prostatectomy, external beam radiation, and brachytherapy. Only variables that were significant on multivariate analysis are included in the nomogram. These clinicians find that socioeconomic status and race factor out of outcome analysis for each stage and grade; these variables are related to diagnosis and time of diagnosis. The nomograms at this point do not incorporate any data on quality-of-life. The outcomes predicted by the nomograms reflect only those drawn from the experience of the physicians at Sloan Kettering.

In the future, the analyses will be used to predict clinical endpoints, such as metastasis-free survival, or to use additional pathology data, such as a percentage of
cores positive and the individual Gleason grades. The current data base is static, meaning that the dataset is analyzed based on a data dump. The data base is currently being revamped with the long-term goal of making the nonogram output dynamic, or based on real-time routine data input. This will involve making the data base Web-based and not disclosing personal health information. The system will be used by other academic centers besides Sloan Kettering in the near future to validate the outcomes observed for the physicians at Sloan Kettering. At MSKCC, it will be used by the departments of urology and medicine. Estimates of the cost to produce these statistical modeling tools for surgery and brachytherapy are $120,000, which does not cover data entry or maintenance costs.

**National Cancer Data Base**

The American College of Surgeons’ National Commission on Cancer (NCC) has established a National Cancer Data Base (NCDB), which is a nationwide oncology outcomes data base for over 1,500 hospitals in the United States. This NCDB is a partnership between the American College of Surgeons and the American Cancer Society to collect and report data on patterns of care and outcomes for all cancers. Annually, the NCDB issues a call for data to approximately 2,000 participating hospitals. The NCDB requests that hospitals voluntarily submit patient information (de-identified) to the national data base and requires hospitals to collect data items required by the Commission on Cancer (CoC). The CoC has constructed a list of specific data elements, data definitions, and coding instructions for the participating hospitals to follow (*Standards of the Commission on Cancer, Volume II: Registry Operations and Data Standards*). The NCDB, while not set up as a longitudinal tracking data base for use by physicians and patients, may provide a platform for recruiting hospitals and physicians across the United States to provide data on prostate cancer patients undergoing oncology treatment.

**Prostate Cancer Research Institute**

Although not specifically a data base used to track longitudinally men diagnosed with prostate cancer, the Prostate Cancer Research Institute (PCRI) is a unique resource for prostate cancer patients who are trying to gather information to make treatment decisions. PCRI was founded by medical oncologists with the goal of patient education to empower men to seek earlier diagnosis and to know the pros and cons of various treatment options. PCRI operates “Patient to Physician” (P2P) via the Internet, which is a moderated mailing list that permits patients to ask focused, clinical questions about
prostate cancer. Patients receive timely responses from physician experts and are encouraged to discuss the information with their doctors. Through this special-interest discussion group, patients can see the responses to questions asked by other patients. This is an example of an informational resource for providing patients with prostate cancer medical information tailored to their unique circumstances, which can be used to engage in shared decision-making with their clinicians.

**Summary**

A number of data base efforts currently demonstrate the feasibility of securing the initial and ongoing participation of patients and physicians in the longitudinal monitoring of outcomes from prostate cancer treatment. These efforts have developed out of an interest in being able to understand the consequences of treatment and to guide future treatment decisions with respect to patient safety and the delivery of high-quality care. Table 1 summarizes the ability of current data base efforts to address the key questions of interest to patients and physicians to help guide treatment choices. As can be seen, there are still important gaps in the information desired as well as how the information is communicated to patients to assist with their decision-making.

None of the data bases reviewed fully realizes the People Like Me concept—one that allows patients, either directly or with their physicians, to create a subset of the data in ways personally relevant to them (by demographic and/or clinical characteristics) and to assess how outcomes of care compare and contrast based on the personal characteristics of the patient. None of the existing efforts has undertaken the step of drawing out the information learned regarding the different probabilities of experiencing different outcomes of interest to patients and creating a decision-support tool that patients can use to assess the risks and benefits of various treatment options. Of the existing data bases, PCOS may represent a viable platform for building a prototype People Like Me data base, because of its attempt to capture a nationally representative cross-section of the prostate cancer population and its focus on evaluating a broad range of patient outcomes.
Table 1: Can the Key Patient Questions be Addressed By Existing Data Bases?

<table>
<thead>
<tr>
<th>Question</th>
<th>PCOS</th>
<th>SEER</th>
<th>CaPSURE</th>
<th>CPDR</th>
<th>Sloan Kettering</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the side effects associated with surgery and how common are they?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>What are the side effects of radiation therapy and how common are they?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>What are the side effects of watchful waiting and how common are they?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>What are the odds that I will be diagnosed with prostate cancer at 50? 60? 70? 80?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Does early initiation of treatment make a difference in clinical outcomes?</td>
<td>No, requires RCT</td>
<td>No</td>
<td>Can evaluate outcomes based on clinical characteristics impacted by stage migration and from date of diagnosis.</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Does early initiation of treatment make a difference in functional outcomes?</td>
<td>No, requires RCT</td>
<td>No</td>
<td>Can provide information on health related quality-of-life (HRQOL) and function based on stage of disease and time from diagnosis and type of treatment.</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>What are the characteristics of patients who are choosing surgery?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>What are the characteristics of patients who are choosing radiation?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes (data base does not represent radiation oncology patients)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>What are the characteristics of patients who are choosing watchful waiting?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Question</td>
<td>PCOS</td>
<td>SEER</td>
<td>CaPSURE</td>
<td>CPDR</td>
<td>Sloan Kettering</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>------</td>
<td>------</td>
<td>-------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>If I get a side effect is there anything I can do? Is it permanent?</td>
<td>Survey asks about side effects. Long-term data analysis will answer question of permanency.</td>
<td>No</td>
<td>Addresses frequency of side effects and impact of impotence and other medications on outcome.</td>
<td>Not sure if data are collected.</td>
<td>No</td>
</tr>
<tr>
<td>Will herbal or supplemental therapies help me?</td>
<td>No. Survey asks if any taken, but does not specify type.</td>
<td>No</td>
<td>Can identify people using complementary and alternative medicine therapies.</td>
<td>Not sure if data are collected.</td>
<td>No</td>
</tr>
<tr>
<td>Does my diet matter?</td>
<td>No</td>
<td>No</td>
<td></td>
<td>Not sure if data are collected.</td>
<td>No</td>
</tr>
<tr>
<td>Does physical exercise make a difference?</td>
<td>No. Survey asks about limitations in activity.</td>
<td>No</td>
<td></td>
<td>Not sure if data are collected.</td>
<td>No</td>
</tr>
<tr>
<td>Is it okay if I drink alcohol?</td>
<td>No</td>
<td>No</td>
<td>No. Has information on smoking and alcohol use, not sufficient to determine a causal effect.</td>
<td>Not sure if data are collected.</td>
<td>No</td>
</tr>
<tr>
<td>Is it okay if I smoke?</td>
<td>No</td>
<td>No</td>
<td>No. Has information on smoking and alcohol use, not sufficient to determine a causal effect.</td>
<td>Not sure if data are collected.</td>
<td>No</td>
</tr>
</tbody>
</table>

**Strategy for Constructing a National Prostate Cancer Outcomes Data Base**

To create a blueprint for constructing a national prostate cancer outcomes database, we define key design aspects of the data base to fully actualize the People Like Me concept. As a caveat, the strategy outlined below will require refinement based on various assumptions about how the data base should be designed and what it should include. For example, we are assuming that data collection commences when a patient has a positive diagnosis of prostate cancer. An alternative, not discussed here, is to commence data collection with PSA screening so as to track patients prior to becoming active cancer cases; however, one concern with this strategy is the potential for bias because not all men elect to undergo screening. Consequently, finding the appropriate at-risk population to participate would be difficult.
Content of the Data Base

The exact content of the data base needs to be determined by the goals of the data base and specific set of questions that it is designed to address. While we have attempted to identify the key questions of interest through our consultations with clinicians and patient advocates, we believe more work is required to refine the list of questions and to achieve consensus by clinicians, researchers, and patients on the outcomes that will be assessed and reported to various audiences. The contents of the data base should not be static and the data base will have to be designed to accommodate and reflect any changes in the treatments available and/or disease course over time. This may require the addition and subtraction of variables as well as changing data definitions. In our discussions with staff from various data base endeavors, it was noted that the data elements contained in the data base require continual review and updating—an important aspect of the ongoing maintenance of any data base effort.

At a minimum, data should be collected on all standard therapeutic options including watchful waiting, radical prostatectomy (surgery to remove the prostate), brachytherapy (radioactive seed implantation), external beam radiation, orchiectomy (removal of the testes), and androgen deprivation. Treatment data need to be detailed to understand the exact type of treatment provided to the patient (e.g., whether treatment involved conventional external beam or conformal external beam radiation). Given significant interest among patients in alternative therapies and virtually no studies to examine outcomes, the People Like Me data base may provide an opportunity to collect information on alternative treatments (either alone or in combination with standard treatments) and to assess outcomes. Should a decision be made to include alternative treatments, substantial effort will need to be applied to defining these treatments so that data can be collected in a standardized way. No existing data structure on alternative treatments has been designed or tested that is ready for immediate application, so considerable work will be required to address this issue.

In addition to the treatment variables, the data base should capture

- contact information for the patient—name, address, phone, e-mail address, consent
- contact information for the provider—name, address, phone, e-mail address
- dates of treatment
- demographic/patient information—age, life expectancy, family history of prostate cancer, history of other cancer, insurance coverage, education,
income, race/ethnicity, primary language spoken, satisfaction with treatment
• clinical information—pretreatment clinical stage (digital rectal exam), pretreatment Gleason score, pretreatment total PSA, comorbidities, use of hormone therapy, pretreatment urinary, sexual, and bowel function, list of visits and details for each visit, hospitalization information, biopsy, medications, pathology, pre- and posttreatment physical, mental and social functioning, posttreatment Gleason scores and PSA
• outcome information—survival (one, five, 10 year), fatigue, posttreatment physical, mental and social functioning, posttreatment urinary, sexual, and bowel function
• provider characteristics (physician and hospital)—board certification, specialty, annual volume of procedures, teaching versus community hospital, and other proxy variables of differences in the care a patient might receive (e.g., urban versus rural).

Sources of Data
Information to populate the data base will need to come from both clinicians and patients. Information such as lab values, pathology reports, imaging, staging of disease, clinical description of the tumor, description of the treatment, complications, functional status, intermediate and long-term outcomes, and characteristics of the physician (e.g., volume of procedures, training) can be compiled from the physician—preferably from clinical records. Patients should not be asked to provide information about treatments and staging, for which they may have poor recall. Patients can provide information on their functional status (physical, mental, and social) and intermediate and long-term outcomes. To facilitate the buy-in of physicians to provide information on a sustained basis, efforts would need to be made to avoid creating redundancies in record-keeping; this could be achieved if the data base served a record-keeping function for their practice.

Approaches to Obtaining Data
The differing methods of recruitment used by existing data bases have different implications for the cost of the project and potential for bias. One strategy is to mimic the PCOS and start by identifying a sample of patients. PCOS draws a sample of patients from the SEER cancer registry data base, which by design is national in scope and seeks to capture a representative sample of cancer patients. Following a design such as PCOS offers the advantage of being able to identify and recruit a representative sample of patients into the data base. The downside of this approach is that physicians
are not directly engaged to work with patients to encourage and reinforce participation, so response rates to surveys are likely to be lower. It also potentially involves engaging a larger number of clinical practice sites from which to collect clinical data on patients because this approach does not require recruiting all prostate cancer patients within a given physician’s practice. Consequently, the fixed costs are likely to be higher in this model if only a small number of patients in any single provider’s practice are part of the data base.

An alternative approach is to sample providers first—both oncologists and urologists—and then to recruit patients from the selected providers. This is a cluster-sample strategy, and the process might first select cities and some rural areas that represent a broad cross-section of patient and clinical practice sites nationally, then sample providers within those geographic areas, and finally take either some or all patients for each provider in the sample. This strategy has the advantages of engaging providers to serve as agents to help in recruitment of patients and may achieve cost economies by gathering multiple patients and their data from each practice site chosen for inclusion in the study. Unlike the provider-focused recruitment efforts used in some of the existing data bases, which were based on convenience samples, we recommend that the recruitment strategy be based on a random or stratified sample of providers who would represent the range of providers that patients with prostate cancer would likely see for treating their cancer. The universe of radiation oncologists and urologists could potentially be identified using Medicare data files.

In any sampling strategy to obtain the data, it will likely be necessary to fill holes in the sample—for instance, for the poor and uninsured or underinsured who may not receive care from the same type of providers identified in the core sample.

**Recruitment of and Participation by Patients and Providers**

Given the positive response among prostate cancer patients at the consultation meetings, it is likely that initial patient participation would be high, provided that the purpose of the data collection was explained, that patients understand that this information would help others with prostate cancer, that the entity compiling the information could be “trusted,” and that patient privacy and confidentiality is protected. Most of the data bases had relatively high rates of initial participation by patients, and those with the highest rates appeared to rely on face-to-face contact with the patient through physician-led recruiting. In PCOS, which relied upon contacting patients via mail or phone interview, response rates to the first request for data were 62 percent of the sampled population.
The greater challenge is ensuring sustained patient participation. Patient attrition was less than 1 percent in the military data base, which involves intensive patient follow up, 5 percent annually in CaPSURE, and 17 percent in the PCOS effort (24-month follow up survey). Some attrition is to be expected because patients die and others view themselves as disease free or “cured” five years after treatment. Engaging in frequent contact with patients appears to improve the likelihood of sustained participation, and most data bases seek to provide incentives to patients to sustain their interest, such as sending them newsletters on prostate cancer.

Physician participation rates were 100 percent in the military data base and 99 percent in the medical abstraction component of the PCOS study. The CaPSURE data base effort also was quite successful in recruiting physician sites, and was limited more by the availability of financial resources to support all sites expressing interest in participating.

**Frequency of Data Collection**

Once the patient is diagnosed with prostate cancer, baseline data collection should commence prior to any treatment (i.e., pretreatment functioning and clinical factors such as stage of disease, PSA, Gleason score, comorbidities). Clinical data are then collected again at time of treatment and follow up visits. In other studies, the frequency of subsequent data collection to evaluate patient functioning—as reported directly by patients—generally occurs at 6 months, 12 months, 18 - 24 months, and then annually thereafter. More-frequent data collection was found to be expensive and burdensome to the patient, and did not produce additional information of value in predicting outcomes.

**Sponsorship**

Sponsorship refers to financial as well as organizational support for the data base project. It is a critical factor in securing the participation of patients and providers. Patients indicated that they view an organization such as the American Cancer Society as a respected and trusted source of patient information and thus a good candidate for sponsoring a data base. Patients also noted that the National Cancer Institute, a federal agency, was a trusted source of information on prostate cancer, and they would view this entity as a credible sponsor. The key concern expressed by patients is that they did not want the sponsor to be an organization that would have access to and share their personal data or gain financially from having access to their data, such as pharmaceutical companies.

The nature of the data collection and products to be derived from the People Like Me prostate cancer data base speak to engaging multiple organizations in sponsoring the effort—including patient advocacy groups like the American Cancer Society Man-to-
Man support groups, Us-Too International, Cap Cure, and the National Prostate Cancer Coalition, research entities such as the National Cancer Institute (both for its research expertise and communications to patients), the American College of Surgeons’ National Commission on Cancer, the Radiation Therapy Oncology Group, and urology and oncology physician organizations such as the American Urological Association and the American Society for Therapeutic Radiation and Oncology. These players should be engaged in helping to define the effort, providing support for patient and provider participation, and potentially as a funding source. A consortium of organizations such as these, in concert with one or more foundations, could potentially provide start-up funds to design and operationalize the data base. Foundations typically support “start-up” efforts, so the project would need to identify long-term financial sponsors to sustain the effort—such as government agencies and/or patient support groups.

**Constructing the Information Tool**

Once the data are compiled for the purposes of determining whether patients experience differences in outcomes from treatment, the next challenge is packaging the information in a way that is usable for consumers and physicians. The approach to organizing and displaying the information needs to be given considerable thought and testing, because the cognitive psychology and decision-making sciences research finds that the way information is framed and the complexity of the data display affect whether the information is used to make decisions and what choices are made (Slovic, 1995; Hibbard et al., 2002). How the material is presented (e.g., mortality versus survival) can influence individual decisions about which treatment to choose; consequently, it is important that the direction and the extent of the influence of the display formats be known and defensible. A key responsibility of the information producer is to produce information that can be used by consumers for choice but will not influence their decision-making in unintended ways. This will require mocking up various prototypes of the decision tool and testing them with patients for design influences—in addition to standard cognitive testing for comprehension and usability. Developers of the consumer information platform will also need to be sensitive to the presentation of complex information for audiences with varying levels of literacy.

As use of the Internet increases and given the structure of Web-based platforms for disseminating information—which allow for easier and cheaper approaches to tailoring information to individual users—the decision aid tool derived from the People Like Me national data base would likely be a computer-based platform. A computer-assisted decision tool can potentially help with the challenge consumers face in processing and weighing various factors to make a choice—especially if the information
is tailored to their own situation. While a computer platform may work for a majority of users, the developers will need to better understand what avenues and preferences consumers have for accessing the information (e.g., print materials, video tapes), because not all patients and their families will have access to the Internet.

The time it takes to develop a computer-aided decision tool will need to be factored into the overall project timeline; this product by itself will likely take two to three years to develop.

**Cost Analysis**

In our review of existing data bases, we asked the staff within each of the projects to define the costs they incurred to develop and operate their data bases on an ongoing basis. While several of the projects were willing to share information about staffing and costs, often these estimates were not of sufficient detail to allow us to construct detailed cost estimates. At best, the cost and staffing figures they did provide give a sense of the magnitude of resource investment that would be required. For several of the projects or for components within projects, we could not obtain cost data because those figures were considered proprietary.

The CaPSURE data base, which tracks 8,100 patients and seeks to add approximately 1,500 new enrollees each year, has a budget of $3.5 million to cover 20 staff at UCSF. The budget covers recruitment of sites, data collection, data cleaning, programming and analysis, research support, and physician incentive payments ($150/patient initially and $100/patient annually thereafter). This figure does not include the costs for staff at the pharmaceutical firm and the third party software vendor that hosts the data collection and reporting tool. Using the $3.5 million estimate, the annual cost per patient is roughly $432.

The PCOS data base has spent between $7 and $10 million over the last eight years (1994 - 2002) to track approximately 3,500 patients, for an average annual cost per patient of $286 (if one assumes that approximately $1 million per year is spent). This budget covers four staff located at the National Cancer Institute and an unknown number of people working with the six SEER registries.

The CPDR data base is funded at $2 million annually and receives additional annual grant support of between $500,000 to $1 million to track 16,000 patients. This works out to an average annual cost of $188 per patient (if one assumes $3 million per year as the budget). This figure covers the CPDR staff and 18 data managers’ salaries. However, it is important to note that this level of funding does not support the full cost of operating the program. All indirect costs (photocopying, office space, travel, phone) are covered by the military, and physicians are prohibited from receiving incentive
payments to participate. If all of these costs were included, the cost of operating the program would be much higher. CPDR estimates that to replicate its data base in the nonmilitary sector would cost approximately $6 million annually.

Final cost estimates will be a function of the size of the patient and physician population tracked, the scope of the effort (number of variables, number of unique sites, amount of data collection, and method of data collection), and whether incentive payments are required to enhance participation. In addition, if a patient-based approach to recruitment is pursued, it may be more costly to implement because of the need to go to many more physician and hospital sites (i.e., the economy of getting all patients at any given physician site is lost and now only a single patient may be enrolled for any given provider).

It should also be emphasized that the cost estimates listed above do not reflect what the actual costs of implementing a People Like Me data base would be, because none of these efforts captures the costs of taking information from the data base and organizing, packaging, and disseminating it to patients and providers throughout the United States. It could cost several million dollars to develop a consumer decision aid; hundreds of thousands of dollars annually to maintain the decision-support tool; and several million more to maintain and disseminate the information. As a result, building a fully functional large-scale data collection system and patient-provider decision support tool could require annual funding in the range of $5 to $10 million.

**A Staged Approach to Building a National Outcomes Data Base**

Establishing the data base will be a multiyear, complex undertaking. Interim steps can be taken to facilitate data base development and also provide useful information to consumers until such time that a national data base is fully up and running and producing information to inform treatment decisions. First steps should include working with physicians and patients to refine the set of questions and specify which clinical and demographic factors are known to be important in influencing the outcomes (i.e., the stratifying variables such as PSA level or Gleason score).

Those who currently operate data bases felt it would be critical to convene a group that would develop a consensus on what outcomes information could or should be released to patients and to clearly define these measures. Another interim step would be to take the set of questions and determine which questions existing data collection efforts or research initiatives have already provided or could provide answers to with additional data mining, and to grade the strength of the evidence. This evaluation would seek to address the questions using the People Like Me concept, by attempting to produce outcomes results by key patient characteristics (age, Gleason score, PSA,
comorbidities). In the process of putting existing data bases to the People Like Me test, we could learn where the information holes are and define appropriate strategies for filling those holes—whether it be a national longitudinal data base or a randomized controlled trial. From the review of the existing data, several researchers thought that information already exists that could be packaged and disseminated to patients to aid them with making treatment choices.

Table 2 outlines the set of steps required to develop a national outcomes data base and the approximate length of time required to accomplish each step. As part of the development process, we include a step focused on exploring opportunities for culling information from existing data sources as a means to flesh out the gaps in the knowledge and data collection processes. However, even if existing data sources provide some of the desired information, there remains the need to invest in analyzing, organizing, and presenting the information in ways that are relevant to patients and that can be used and understood by patients for the purposes of evaluating treatment options. As Table 2 shows, it would take approximately four to five years to fully design and implement the data collection process and to disseminate information to patients.

**Table 2: Timeline and Key Milestones for Developing Data Base**

<table>
<thead>
<tr>
<th>Key Milestones</th>
<th>Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I. Planning Stage</strong></td>
<td></td>
</tr>
<tr>
<td>Define specific purpose of data base, who will use information, and how that information will be disseminated.</td>
<td>Months 1 - 3</td>
</tr>
<tr>
<td>Identify key constituencies to be collaborators in the project (e.g., American Cancer Society, National Cancer Institute, prostate cancer patient advocacy groups, urologists and radiation oncologists). Solicit feedback, input, and support from appropriate patient and provider constituency groups; begin to lay ground work for working collaboratively with these patient and provider groups to promote the benefits of participating in the data base and securing the commitment of physicians and patients to supply data. Form advisory panel to provide policy and technical guidance on the development of the data base.</td>
<td>Months 1 - 6</td>
</tr>
<tr>
<td>Refine the set of questions the data base will seek to answer and engage patients, clinicians, and researchers in a consensus development exercise to specify and define the set of outcome measures that will be measured and reported out to patients and physicians.</td>
<td>Months 1 - 12</td>
</tr>
<tr>
<td>Define and gain consensus from the patient, clinical, and research community on the set of stratifying variables (covariates) that are predictive factors in determining outcomes and that patients will be able to use to query the data base to get answers for People Like Me. Patients with prostate cancer should be asked to specify how they would like to query the data.</td>
<td>Months 1 - 12</td>
</tr>
</tbody>
</table>
Table 2 – Continued

<table>
<thead>
<tr>
<th>Key Milestones</th>
<th>Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review existing data collection efforts and the published literature from these</td>
<td>Months 1 - 18</td>
</tr>
<tr>
<td>studies/project to determine which questions can be answered already by the</td>
<td></td>
</tr>
<tr>
<td>relevant People Like Me subsetting variables. Grade the level of evidence</td>
<td></td>
</tr>
<tr>
<td>contained in the research studies. Identify where the gaps remain that a new</td>
<td></td>
</tr>
<tr>
<td>data base effort could fill. Seek to construct a collaborative pooling of data</td>
<td></td>
</tr>
<tr>
<td>across data collection efforts so as to be able to conduct cross-site analyses,</td>
<td></td>
</tr>
<tr>
<td>and to standardize data definitions, data collection tools, and outcomes of</td>
<td></td>
</tr>
<tr>
<td>interest.</td>
<td></td>
</tr>
<tr>
<td>Define the mechanisms for collecting the data (survey, administrative data,</td>
<td>Months 12 - 24</td>
</tr>
<tr>
<td>chart abstraction).</td>
<td></td>
</tr>
<tr>
<td>Define the sampling method for selecting physicians and patients.</td>
<td>Months 12 - 18</td>
</tr>
<tr>
<td>Take the “key questions” to be addressed by the data base and define the set</td>
<td>Months 12 - 24</td>
</tr>
<tr>
<td>of data elements required and operationalize the definitions. Identify the</td>
<td></td>
</tr>
<tr>
<td>data source and frequency of collection for each element. Determine whether</td>
<td></td>
</tr>
<tr>
<td>new data collection tools are required or whether existing tools can be</td>
<td></td>
</tr>
<tr>
<td>modified.</td>
<td></td>
</tr>
<tr>
<td>Develop new data collection tools (if necessary).</td>
<td>Months 12 - 24</td>
</tr>
<tr>
<td>Prepare training materials (training manual, data code book).</td>
<td>Months 18 - 24</td>
</tr>
<tr>
<td>Design a pilot test of the data collection tools and processing protocols.</td>
<td>Months 18 - 24</td>
</tr>
<tr>
<td><strong>II. Pilot Stage</strong></td>
<td></td>
</tr>
<tr>
<td>Identify a small number of practice sites (&lt;=5) to pilot test the data collection</td>
<td>Months 18 - 24</td>
</tr>
<tr>
<td>methodology.</td>
<td></td>
</tr>
<tr>
<td>Train pilot site personnel on data collection methods.</td>
<td>Months 24 - 25</td>
</tr>
<tr>
<td>Commence data collection in pilot sites.</td>
<td>Months 25 - 36</td>
</tr>
<tr>
<td>Interview staff at each of the practice sites to identify what worked and did</td>
<td>Months 30 - 36</td>
</tr>
<tr>
<td>not work during “implementation.”</td>
<td></td>
</tr>
<tr>
<td>Review data that are submitted during pilot phase to assess completeness,</td>
<td>Months 26 - 36</td>
</tr>
<tr>
<td>consistency, and accuracy.</td>
<td></td>
</tr>
<tr>
<td>Revise data collection tools and procedures.</td>
<td>Months 36 - 37</td>
</tr>
<tr>
<td><strong>III. Full-Scale Implementation of Data Collection</strong></td>
<td></td>
</tr>
<tr>
<td>Expand data collection to more practice sites—recruit providers and patients</td>
<td>Months 38 - 54</td>
</tr>
<tr>
<td>into the study.</td>
<td></td>
</tr>
<tr>
<td>Conduct training of personnel in each site on the data collection.</td>
<td>Months 38 - 54</td>
</tr>
<tr>
<td>Begin collecting data; technical assistance to sites.</td>
<td>Months 39 +</td>
</tr>
<tr>
<td>Ongoing recruitment of sites, training, data collection.</td>
<td>Months 40 +</td>
</tr>
<tr>
<td><strong>IV. Data Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Data processing protocols—data cleaning, auditing, analysis.</td>
<td>Months 39 +</td>
</tr>
<tr>
<td>Organize results for distribution to providers and patients.</td>
<td>Months 48 +</td>
</tr>
</tbody>
</table>
Table 2 – Continued

<table>
<thead>
<tr>
<th>Key Milestones</th>
<th>Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>V. Development of Information/Decision Support Tool</strong></td>
<td></td>
</tr>
<tr>
<td>Meet with patient representatives to define the content of the information</td>
<td>Months 12 - 15</td>
</tr>
<tr>
<td>resource and how it should be communicated.</td>
<td></td>
</tr>
<tr>
<td>Meet with physician representatives to define the information they would like</td>
<td>Months 15 - 18</td>
</tr>
<tr>
<td>to see presented for use by physicians to assist with making treatment</td>
<td></td>
</tr>
<tr>
<td>decisions.</td>
<td></td>
</tr>
<tr>
<td>Review the cognitive literature on consumer/patient decision-making to</td>
<td>Months 15 - 20</td>
</tr>
<tr>
<td>guide the design of the information resource.</td>
<td></td>
</tr>
<tr>
<td>Develop a mock-up of a prototype decision support tool for review by patients</td>
<td>Months 20 - 26</td>
</tr>
<tr>
<td>and physicians.</td>
<td></td>
</tr>
<tr>
<td>Conduct testing with patients/consumers for design influences (e.g., framing,</td>
<td>Months 26 - 30</td>
</tr>
<tr>
<td>use of language), comprehension, and usability.</td>
<td></td>
</tr>
<tr>
<td>Build consumer information tool.</td>
<td>Months 36 - 48</td>
</tr>
<tr>
<td><strong>VI. Dissemination/Public Awareness</strong></td>
<td>Months 48 - 56</td>
</tr>
<tr>
<td>Build awareness among the patient and provider community that a tool is</td>
<td></td>
</tr>
<tr>
<td>available to assist with making treatment decisions (public education</td>
<td></td>
</tr>
<tr>
<td>campaign).</td>
<td></td>
</tr>
</tbody>
</table>

Feasibility Considerations: Issues That Need to be Addressed in Building a National Prostate Cancer Outcomes Data Base

_Bias in the Sample of Patients_

Observational data bases, properly designed, should be inclusive of all patients who have the condition and receive treatment, even if it is watchful waiting. How physicians and patients are recruited into the longitudinal tracking effort will influence the degree to which one can make outcome predictions that are valid and generalizable to the broad population of prostate cancer patients. The results of most existing prostate cancer data collection efforts may be biased or of limited generalizability because they only include patients who are on specific treatment protocols, who are from medical centers with selected characteristics, or who are in certain age groups. For the People Like Me outcomes data base that would be used by patients and providers to compare treatment options, the strategy for recruiting and including patients in the data base should allow for the collection of data from patients and physicians that represents the full patient population and set of treatment options (surgery, radiation, hormone, watchful waiting). Doing so will minimize the potential for bias in the results.

The initial sampling or recruitment strategy—if provider based—should focus on getting a broad mix of providers (e.g., urologists, radiation oncologists, and primary care providers) and ensuring the highest possible patient participation in each of those sites.
A patient-based recruitment strategy, depending on the level of response rates achieved, runs the risk of nonrespondent bias if response rates are low. Consequently, efforts (i.e., aggressive follow up, provision of incentives) should be made to maximize response rates and also to assess the degree to which bias may be present. One approach is to assess whether the sample of participants is different from nonparticipants, which requires the collection of a small number of descriptive data elements on all potential participants (i.e., those invited to participate). Many of the existing data base efforts reviewed for this report do attempt to collect a minimum set of data elements on all persons invited to join the data base.

Size of the Patient Sample Needed to Develop an Outcomes Data Base

The sample size required for an outcomes data base is mainly a function of the number of People Like Me patient characteristics (e.g., age, race/ethnicity, stage of disease) or subsetting variables desired that are related to the outcomes of interest. In a study of prostate cancer outcomes by Litwin et al. (2000), the authors identify a handful of clinical covariates that might be important covariates associated with the various prostate cancer outcomes. These variables include patient age, patient life expectancy, pretreatment total PSA, clinical stage, Gleason grade, family history of prostate cancer, history of other cancer, and comorbidity indicators. Additional covariates would likely include treatment type, the use of hormone therapy, insurance coverage, race/ethnicity, education, and income. As more subsetting variables are added to allow patients and their physicians to tailor the output in a way unique to the individual patient, more patients must be tracked to ensure adequate sample sizes in each analytic cell (e.g., an analytic cell could be defined by a specific age range of the patient, race/ethnicity, and treatment type).

It is more difficult to detect differences among small subgroups in the population than among larger groups. Estimates for small subgroups of patients will have larger standard errors (i.e., more uncertainty in the estimate). Consequently, it may be worth considering, for certain variables such as age, displaying information across a range (e.g., ages 50 - 60) to address the problem of small sample sizes. Clinicians and patients would need to jointly determine acceptable and biologically appropriate ranges for variables that would represent aggregations, as well as the subsetting variables and outcomes of interest.

Because most patients would like information to be closely tailored to their own experience, it will be necessary to collect data on a large number of people across a diversity of patients. This desire will need to be balanced with cost issues associated with tracking more individuals and the prevalence of selected outcomes. For example,
to study rare treatment combinations or rare outcomes, it would be essential to get data on all patients (as in a census or registry like the End Stage Renal Disease program)—something that would be more difficult and costly than sampling patients.

To estimate the size of the sample needed to develop an outcomes data base requires making certain assumptions. For the calculations below, we have assumed that the outcomes data base will be used for regression analyses that are multivariate in nature. This includes contrasts between two groups (or t-tests) as a special case. However, it excludes reducing the data by conditioning on certain variables or strata and then contrasting outcomes within the subset of the data.

We have also assumed that, in most cases, higher-order interactions beyond two-level interactions are not significant or policy relevant (i.e., for large data sets interaction terms between two variables may be statistically significant but the coefficients are too small to matter in practice). Because of the normal approximation to the binomial distribution we can use power calculations based on the normal distribution with variance 0.5. Table 3 shows sample sizes needed under various scenarios. We have computed the estimates as a function of the smallest detectable difference that one would expect to observe between outcomes of interest (e.g., a 20-percentage-point difference in mortality) and the possible presence of two-level interactions. Calculations are based on a significance level of $\alpha = 0.05$ and 80 percent power to detect the effect. The calculations represent the worst-case situation where the probability is $p_1 = 0.5$.

Table 3: Sample Size for Each Category

<table>
<thead>
<tr>
<th>Number of patients for each cell created by subsetting</th>
<th>Smallest detectable difference in outcomes (percentage point)</th>
<th>Presence of two-level interaction effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>99</td>
<td>20</td>
<td>No</td>
</tr>
<tr>
<td>197</td>
<td>20</td>
<td>Yes</td>
</tr>
<tr>
<td>175</td>
<td>15</td>
<td>No</td>
</tr>
<tr>
<td>349</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>393</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>785</td>
<td>10</td>
<td>Yes</td>
</tr>
<tr>
<td>1,570</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>3,139</td>
<td>5</td>
<td>Yes</td>
</tr>
</tbody>
</table>

If the subsetting variable or covariate—such as race—contains more than two levels, the sample size shown in Table 3 above is required for each of the levels. If the covariate is continuous these calculations still hold and can be considered conservative. If the outcome variable is continuous, the approach is still valid, but Table 3 does not
apply because it assumes a percentage for the smallest detectable difference. Instead, an estimate of mean and variance of the outcome variable is needed to do the power calculations.

To help illustrate the implications of Table 3, we estimate the required number of patients for the following data base example. Let’s say we are interested in evaluating differences in five-year survival rates for different forms of treatment on the following patient characteristics (variables with the number of “levels” shown in parentheses behind each): age (continuous), pretreatment PSA (continuous), clinical stage (four levels), Gleason grade (continuous), comorbidities (three levels), race/ethnicity (five levels), education (three levels), and insurance coverage (three levels). Assuming that the data base is perfectly balanced across covariates and that we are not throwing away information by subsetting (i.e., we are producing estimates using regression analysis that uses all of the information available), then we would take our largest number of levels—in this case five for race/ethnicity—and multiply by the “number of patients for each cell” shown in Table 3. Thus, for a 10-percentage point difference in five-year survival and no interaction effects (393 patients per cell), we would minimally need 1,965 patients. For a five-percentage point difference in outcome with no interactions (1,570 patients per cell), we would need 7,850 patients. If the variable is not balanced, more patients would be required.

Table 3 represents a worst-case scenario where the probabilities of interest are around 0.5. Table 4 considers several small probabilities explicitly, and the resulting sample sizes are much smaller (roughly one-third to one-half the previous sample sizes). Table 4 stems from ordinary power calculations based on a binomial distribution. Interactions were not considered. Calculations are based on a significance level of $a = 0.05$ and 80-percent power to detect an effect.

**Table 4: Sample Size for Each Category as a Function of Various Smaller Probabilities**

<table>
<thead>
<tr>
<th>Number of patients for each subsetting variable</th>
<th>p1</th>
<th>p2</th>
<th>Smallest detectable difference (percentage point)</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>0.05</td>
<td>0.25</td>
<td>20</td>
</tr>
<tr>
<td>88</td>
<td>0.05</td>
<td>0.20</td>
<td>15</td>
</tr>
<tr>
<td>160</td>
<td>0.05</td>
<td>0.15</td>
<td>10</td>
</tr>
<tr>
<td>474</td>
<td>0.05</td>
<td>0.10</td>
<td>5</td>
</tr>
</tbody>
</table>
We provide another example to illustrate the sample size calculations. In prostate cancer, one of the outcome variables of interest is presence or absence of hospitalization, medical or surgical treatment for cystitis, proctitis, hematuria, or rectal bleeding. Each of these is a dichotomous variable. Suppose we determine that the covariates to be used in the regression analysis are patient has/does not have insurance, pretreatment total PSA, and clinical stage. Further suppose that we are interested in all two-level interactions and that three-level interactions are negligible.

To detect a difference in the outcome variable of 10 percentage points (e.g. 50 percent versus 60 percent) under the usual power and significance assumptions, we would require 785 patients in each variable category. Thus, there need to be 785 patients with insurance and 785 patients without insurance. There also needs to be 2 x 785 patients whose PSA score roughly covers the range of interest (i.e., is not concentrated in a small area of the range). If the data base is perfectly balanced across all covariates this can be accomplished with 2 x 785 = 1,570 patients. In practice, the number will be somewhat larger because the data base is unlikely to be perfectly balanced across all categories. If we consider more than three explanatory variables, the sample size requirement does not change. However, the more variables there are, the more likely it is that some categories do not have 785 patients. Since the same number of patients is needed for each category the sample will tend to be larger.

Some variables like race have more than two levels. The number of patients required, as determined from Table 3, needs to be available from each level within the variable category (e.g., for race, 785 whites, 785 African-American, 785 Asian, and so forth). Suppose it is known that African-Americans are difficult to recruit and that the data base may only successfully recruit 200. Based on Table 3, a sample of 200 African-American patients means that we would only be able to detect a difference of 20 percentage points for outcomes with interaction effects and a difference of (at least) 15 percentage points for main effects.

Sample size calculations always presume that the sample is drawn at random. Outcome data bases are generally not random because it is highly impractical and very costly to sample patients at random. To do so would require the existence of a national registry of patients. In theory, the lack of randomness limits statistical inferences that can be drawn from the sample. The lack of randomness is somewhat alleviated by ensuring that the sample is representative with respect to the relevant variables (e.g., all treatments, all treatment settings, and types of providers). Because the People Like Me subsetting approach explicitly considers important variables as covariates, it increases
the likelihood of having a group of patients that is representative of the population of interest.

Throughout the sample size calculations shown thus far, we have assumed that the outcome variable is dichotomous (e.g., yes or no). If the outcome is continuous, the sample size needed will be substantially lower. This implies that the easiest way to increase power in analysis is to avoid creating discrete variables (either outcomes or predictors) wherever possible. For example, instead of collecting information on whether or not the patient survived until five years after a treatment (yes or no), it would be preferable to record the length of survival and conduct a survival analysis.

The sample sizes proposed in Table 3 are considered to be conservative. Throughout, we have assumed that we are trying to detect a difference of 0.5 versus an alternate value. This is conservative because, for example, it is much easier to detect 0.05 versus 0.1 than 0.5 versus 0.55. Therefore, the sample sizes displayed in Table 4 are much smaller.

The “curse of dimensionality” refers to the following phenomenon: the more variables that are considered simultaneously, the sparser the data appear. This “curse” applies to the following situation: If I wish to understand the outcome under different treatment alternatives for “people who are exactly like me” (i.e., for people who match my characteristics on all covariates), the number of people I am comparing to will be very small. The ability to draw conclusion on such a subgroup will be limited. A regression approach circumvents the curse of dimensionality by making use of the assumption that no higher-order interactions are relevant. This holds true in most practical situations.

As noted above, the desired patient sample size is dependent on whether one wishes to estimate interaction terms (to account for nonlinearities in the data) with sufficient power. The sample size recommendations accommodate both multivariate and univariate analyses. We do not accommodate within-strata analyses. (Note: A stratified or univariate display of the data may be misleading to patients because the outcomes represented do not take into account [or control for] other variables that may be important factors in determining the outcome. For example, looking at outcome results by age alone will miss the effect that clinical stage of the cancer has on predicting the outcome of interest, such as likelihood of survival.)

**Outcomes to Be Measured**

At the outset of the project, it will be critical to specify the key outcomes (e.g., survival, functioning, complications) that should be tracked. The process will benefit from soliciting input from and working toward consensus with the patient and clinical
community that the specified outcomes are in fact the relevant endpoints of interest to the end users of the data base. Once the outcome measures are agreed to, additional work will be required to clearly define the measures (creating operational definitions) for the purposes of uniform data collection. For example, the measure would need to define the starting point (baseline could be from diagnosis or treatment) and the end point (e.g., in-hospital mortality, 30-day mortality, five-year mortality).

Throughout the design and ongoing operation of the data base, it will be necessary to solicit clinical input into what variables are being collected, how are they defined, and what role they play in analysis (are they predictors, are they descriptive). As occurs with any data collection effort, the data base will evolve over time—both its contents and its functions. Therefore, it will be important to seek ongoing feedback from data entry personnel in clinical practices and hospitals, clinicians, researchers, and patients to improve the data and the data base quality.

**Data Base Architecture**

One piece of feedback received from the existing data base projects was the need to store “raw” values to ensure flexibility to derive variables for various analyses. Because key variable definitions may change over time (e.g., the determination of clinical staging), having access to the raw data elements used to derive variables allows flexibility over time to use information collected in earlier time periods for analyses as definitions change. In examining all data elements, it will be important to determine which can be broken down into more granular elements and to be sure to track information at this level. A relational data base architecture will likely be most useful to allow the tracking of many different pieces of information that can be linked.

**Data Quality**

One of the areas that will affect the ability to produce reliable and valid outcome results is the underlying quality of the data supplied by clinicians, hospitals, and patients. As always, there is a risk of variation in how data are captured across sites that has nothing to do with real variation in patients, their care, or their outcomes. An important step in reducing variation is to have clearly defined and agreed-to variables and definitions for those variables. Other strategies to reduce variation in coding across sites is to have standardized data collection instruments and to engage staff from the sites providing the data in regular training sessions. However, despite these efforts, the project staff operating the data base will need to develop programs (i.e., data quality reports looking for out-of-range values and inconsistent information, occurrences of values for data elements that are well outside the average observed across all institutions) and maintain oversight to catch and correct problems with coding. These
efforts can be reinforced with periodic audits of data being submitted by providers to verify the accuracy of data submissions. Auditing can be a costly activity because it usually involves going to a practice site and abstracting information from a clinical record. Audits can be designed to be cost-effective by focusing on a limited number of data elements (e.g., those that are most important), a limited number of sites each year, and being purposeful in selecting sites for audit (e.g., those that appear to have the most difficulty with coding or seem to have statistically better or worse performance results).

Data Output

There will likely be multiple end users of the data, including researchers, clinicians, and—most important—patients. Supporting the needs of the various end users will require understanding what information they would like to have and how they propose to manipulate the information. Patients might be given some flexibility to customize their searches based on a handful of clinical and demographic characteristics, whereas clinicians and research staff might be provided broader access to de-identified information for running analyses. Customization of output allows for different uses of the information by end users, but it is also more expensive to maintain. For patients in particular, little work has been done on how to develop an effective user interface for presenting outcomes information. Developing the interface will require an evaluation of alternative modes for efficient presentation of complex data to individuals with diverse information needs and preferences for involvement with treatment decisions.

Protection of Confidential Information and Patient Privacy

Any time someone gives personal information to someone else, maintaining the confidentiality of that information and ensuring patient privacy is a key concern. For the purposes of creating a longitudinal outcomes data base on prostate cancer, we would need to collect contact information on both the patient and the provider (including name, address, and possibly telephone number) for follow up data collection (both clinical and survey information). To gain the cooperation of patients and providers (and their affiliated institutions), it will be necessary to develop a plan to protect human subjects and have a policy for data safeguarding that is reviewed and approved by an Institutional Review Board.

Each patient would need to be assigned a research ID number when he completes the initial questionnaire. This way, information regarding functional status and outcomes can be kept separate from contact information, ensuring that people accessing the data base cannot obtain contact information for contributors. Patients must be assured that all data will remain confidential and that others will only be able to access
data in aggregate form. Prior to the collection of any data, a procedure would be required to ensure that patient and physician privacy concerns are addressed.