SPECIAL REPORT

The Marshfield Clinic Personalized Medicine Research Project: 2008 scientific update and lessons learned in the first 6 years

Catherine A McCarty[1†], Peggy Peissig[2], Michael D Caldwell[3] & Russell A Wilke[4,5,6]

[†]Author for correspondence
[1]Center for Human Genetics, Marshfield Clinic Research Foundation, 1000 N Oak Avenue (ML1), Marshfield, WI 54449, USA
Tel.: +1 715 389 3120;
Fax: +1 715 389 4950;
E-mail: mccarty.catherine@mcrf.mfldclin.edu
[2]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, 1000 N Oak Avenue (ML1), Marshfield, WI 54449, USA
[3]Department of Surgery, Marshfield Clinic, 1000 N Oak Avenue (ML1), Marshfield, WI 54449, USA
[4]Department of Medicine, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee,WI, USA
[5]Department of Pharmacology and Toxicology, Medical College of Wisconsin, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee,WI, USA
[6]Human and Molecular Genetics Center, Medical College of Wisconsin, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee,WI, USA

The Marshfield Clinic Personalized Medicine Research Project is the largest population-based biobank in the USA, with the ability to recontact subjects to obtain additional information to facilitate gene–environment studies. Nearly 20,000 adults have enrolled in the Personalized Medicine Research Project since 2001, after providing active written consent to access their Marshfield Clinic medical records to define phenotype and providing blood samples from which DNA, plasma and serum samples were stored. Numerous studies are underway in the area of pharmacogenetics and genetic epidemiology. In addition to the scientific discoveries being made, much has been learned regarding biobanking and the management of large amounts of data being generated. The purpose of this paper is to share the advice provided by the external Scientific Advisory Board and the scientific lessons learned along the way to build this research infrastructure and facilitate its use.

Keywords: biobank, genetic epidemiology, methods, pharmacogenetics

The Marshfield Clinic Personalized Medicine Research Project (PMRP) was launched, along with other international biobanking efforts, in 2002 [1]. This was followed by a Grand Challenge issued in 2003 by Francis Collins and his colleagues from the National Human Genome Research Institute (MD, USA) to 'develop robust strategies for identifying the genetic contributions to disease and drug response' [2]. In this visionary paper, Collins *et al.* mentioned the PMRP and two other large biobanks as longitudinal population-based cohort studies that could help to address this Grand Challenge.

The ultimate goal of the PMRP is to translate genetic data into specific knowledge about disease that is clinically relevant and will enhance patient care, with the short-term goal of establishing a database to allow research in genetic epidemiology, pharmacogenetics and population genetics. The PMRP is a biobank with DNA, plasma and serum samples and access to the Marshfield Clinic medical records of nearly 20,000 actively consented adults aged 18 years and older. Details about the study methodology have been published previously [3]. The purpose of the current review is to provide a scientific update on the PMRP after 6 years of enrollment, to provide information regarding projects currently using the PMRP database, and to survey some of the lessons learned during the establishment of this biobank for studies involving genetic epidemiology and pharmacogenetics.

Scientific Advisory Board

A tremendous amount of effort was invested in the preparation of the PMRP biobank, long before its activation in 2002. The details of our efforts have been presented previously within *Personalized Medicine*, within the specific context of discussing infrastructure development (Scientific Advisory Board [SAB], Community Advisory Group, and the Ethics and Security Advisory Board) [3]. The SAB met twice prior to the launch of the PMRP in September 2002, and once again 15 months after enrollment had commenced. Tables 1–3 summarize the recommendations made by the SAB at their three meetings and the actions taken. Funding to support the SAB ran out before the cohort was large enough to undertake studies using the biobank.

Oversight committee

Partly in response to the advice of the SAB that a committee be formed to review and approve applications to use the biobank, Tissue Access Guidelines and an Oversight Committee were written and formed. The Tissue Access Guidelines are available on the PMRP website [101]. In brief, in addition to Marshfield Clinic Institutional Review Board approval, scientific merit and approval by the Oversight Committee is required for the release of samples from the biobank. Where external scientific merit has not been approved through a traditional peer-review mechanism, such as the NIH for funding, the

10.2217/17410541.5.5.529 © 2008 Future Medicine Ltd ISSN 1741-0541

Table 1. Recommendations from the first Scientific Advisory Board meeting, held in November 2001, and actions taken.

Recommendation	Actions taken
Collect the DNA as planned	Recruitment commenced in September 2001
Include a control population	Cohort is population-based, allowing for case–control or case–cohort designs
Limit yourself to a small number of focused studies, perhaps two to three	Initially, the warfarin pharmacogenetics study was the only study being undertaken
Create a process for people to develop project proposals that include the core ideas about hypothesis, preliminary data, how many subjects, study power, and how the phenotyping will be undertaken, in 3–5 pages	The Research Foundation already has a process to review studies for scientific merit. We are looking to streamline the process for PMRP studies where institutional funding is not requested
Consider the warfarin pharmacogenetics study as a proof-of-concept study where you could have results very quickly and demonstrate the success if you assign sufficient resources to the project	The warfarin pharmacogenetics study became the main study undertaken. To speed the process, recruitment was performed outside the biobank because the biobank was conducting general population recruitment
Identify external collaborators – academic versus industry	The decision was made to concentrate on academic partnerships initially, in part because of a negative (although not universal) reaction from the Community Advisory Group to discussion regarding potential collaboration with industry
Consider allowing de-identified DNA samples to be sent off site for genotyping	The written informed consent document allows for sharing of samples and data. Sharing of both data and samples has already occurred

PMRP: Personalized Medicine Research Project.

Research Committee at the Marshfield Clinic will review research proposals. The composition of the Oversight Committee is as follows:

- Director of Medical Research, Chair of Oversight Committee
- Principal Investigator of the PMRP
- Department Chair of Medical Genetics
- Director of the Center for Human Genetics
- Director of the Marshfield Clinic Laboratories, or his/her designee
- Director of Informatics and Decision Support
- One or two members from the University of Wisconsin, Madison (WI, USA), ideally one with expertise in statistical genetics

The Director of Medical Research has discretion to add additional *ad hoc*, permanent or *ex-officio* members as necessary. The Oversight Committee meets monthly to review applications to use the biobank and to prioritize infrastructure projects, such as the many information systems projects to facilitate access and quality assurance for the three specimen types. They do not assess scientific merit; rather, they consider issues such as the amount of sample remaining and the quantity of sample requested for a study.

Requests to the Oversight Committee to access PMRP samples are submitted and tracked online. Laboratory staff do not release samples until all approvals have been documented, including a Material Transfer Agreement if necessary. The legal department at the Marshfield Clinic developed a template for Material Transfer Agreements involving PMRP samples to speed up the approval process.

So far, no requests to access the plasma or serum samples have been reviewed by the Oversight Committee. We anticipate that the decisions surrounding the release of those specimens may be more challenging, given the smaller quantity and the lack of an equivalent technology to whole-genome amplification to allow preservation of the original sample.

Sample tracking

A significant challenge for the PMRP was the ability to manage the collection and tracking of biological samples. The PMRP uses Marshfield Clinic's practice management and laboratory information systems (LIS) for subject recruitment and DNA, plasma and serum sample collection. The initial decision to use these systems was based on cost and the length of time it would take to implement another information system to support the recruiting and sample collection efforts. The LIS was not designed to link multiple samples together for an individual or to track the location of an individual's genetic material within the freezer or elsewhere, after several generations of plating. In addition,

Table 2. Recommendations from the second Scientific Advisory Board meeting, held in April 2002, and actions taken.

Recommendation	Actions taken
Publish a paper in a peer-reviewed journal defining the characteristics of the population and demonstrating the power of the population, incorporating genetic data if it did not slow the time to publication	Two papers were published, one in 2005 [3] and one in 2007 [18], describing the biobank and available data
Expand on the collection of family cohorts	Information about relatedness is collected on the enrollment questionnaire. The decision was made not to recruit families specifically but to build a population-based biobank for association studies
Proceed with clinical studies; complete the warfarin study within 6 months	The initial paper from this study was published in 2004 [27]; it was delayed, in part, by recruitment difficulties in the clinic. The prospective clinical study commenced immediately after the retrospective study was completed, with results published in 2005 [29]
Consider alliances with other entities to obtain collections of SNPs/haplotypes and software for analysis	Collaborations have been initiated with scientists at the University of Wisconsin (WI, USA), with support from the Clinical and Translational Science Award, and with scientists at Vanderbilt University (TN, USA)
Continue diligence in seeking appropriate methods for genotyping	Genotyping technologies are continually assessed. In the context of multicenter studies, whole-genome association genotyping is being conducted at core laboratory facilities. Candidate-gene studies in-house are completed on the Sequenom® platform
Consider how in-house SNP discovery will be undertaken	This has not been undertaken due to limited staff and the decision to prioritize the development of the phenotypic infrastructure

quality information for each sample was stored in another unrelated database and needed to be merged into a single system for efficient specimen management. Owing to these limitations, the PMRP evaluated options to either modify the existing LIS for biospecimen management or purchase a commercial system. At an early stage, focus was directed at evaluating commercial biospecimen tracking systems because the internal software development resources were limited. Four commercial systems were evaluated, with a final selection of the Nautilus® biospecimen tracking system (Thermo Fischer, PA, USA). The Nautilus product offered a complete specimen tracking system with many out-of-the-box customizable features and integration tools that would allow PMRP software developers to customize user interfaces and interface to Marshfield Clinic information systems. The laboratory technicians found the Nautilus user interface easy to follow and felt that the solution met the functionality requested. The PMRP laboratory and informatics teams are currently redesigning many workflow processes and integrating Nautilus to promote efficient specimen handling for PMRP and other research activities.

Informatics has been a critical part of the PMRP since its infancy. One of the basic goals of this project was to use both genetic sequence data and existing clinical data to accelerate the study of diseases. To this end, PMRP capitalizes on the use of Marshfield's in-house developed Cattail's Software Suite to provide most of the clinical data for this effort. Data from the Cattails Suite are deposited into a centralized data repository, where they are used for the ongoing clinical care of patients. Data from the centralized data repository are transferred nightly to Marshfield Clinic's data warehouse, where the data are cleansed and integrated into the existing data warehouse structure. Prior to adding data sources to the data warehouse, the data source is modeled and then integrated into the enterprise data model. This enables the data warehouse development team to standardize names, data types and variable definitions. The data are also grouped into logical subject areas that are understandable to the users of the data. Internally developed applications and commercial bioinformatics tools provide access to the data for a variety of clinical and business users. Figure 1 denotes the strategy used to develop the data warehouse. Within the data warehouse, patient data from as early as 1960 are available electronically and linked together by a unique patient identifier. The data carry a time stamp so historical profiles of clinical activity can be developed for each PMRP participant. Data from the data warehouse are used for PMRP studies and enable scientists to accurately identify and categorize patient's phenotypes for research studies.

Table 3. Recommendations from the third Scientific Advisory Board meeting, held in November 2003, and actions taken.

Recommendation	Actions taken
Have a narrow focus on a smaller set of projects with more depth	After initiating three internal studies initially, the decision was made to further encourage external collaborations and use of the database, thus the number and scope of projects increased
Define an overall concept, vision and well-defined priorities. Consider hiring a Scientific Director	The overall goal of the PMRP is to translate genetic data into specific knowledge regarding disease that is clinically relevant and will enhance patient care, with a short-term goal to establish a database to allow research in genetic epidemiology, pharmacogenetics and population genetics. Funding limitations have precluded the hiring of a full-time Scientific Director, but the Oversight Committee fulfills this role
Develop more external collaborations	Members of the Pharmacogenetics Research Network were invited to Marshfield in July 2006 to learn more about the PMRP. This visit resulted in the establishment of several collaborations, as mentioned in **Tables 5 & 6**. The Clinical and Translational Science Award is helping to facilitate increased collaboration with scientists at the University of Wisconsin (WI, USA)
Decide on the relative balance between creating the PMRP as a national resource versus a research activity at Marshfield. If it is going to be a national resource, more details about data sharing need to be available	The decision was made to further encourage external collaborators, in part to realize the potential of the resource for genetic discovery. Guidelines for access are available on the PMRP website [101]
The SAB should meet more regularly, perhaps twice per year	The grant that supported the SAB meetings ended, so additional meetings could not be held

PMRP: Personalized Medicine Research Project; SAB: Scientific Advisory Board.

Personalized Medicine Research Project participants are asked to supply information on their health habits, environment, family history and parent–sibling–child relationships. These data reside in the data warehouse and are thereby useful for PMRP studies. The Institutional Review Board and PMRP Oversight Committee provide approval for the use of these data. Data security controls have also been developed to ensure use of these data for PMRP studies only unless appropriate approval has been granted for additional use.

Figure 2 denotes three distinct sources of data used by the PMRP. The data within this figure flow from left to right into the Personalized Medicine Research Database (PMRD). The clinical data flow was described previously in the data warehouse discussion. Clinical data consisting of diagnoses, laboratory values, procedures, insurance claims, clinical registry, billing and medications data, found within the data warehouse, are used to phenotype PMRP subjects. Once a study population (and associated study data) are identified and validated, data analysis variables are packaged and sent to an encryption process prior to being merged with the genetic data located in the PMRD.

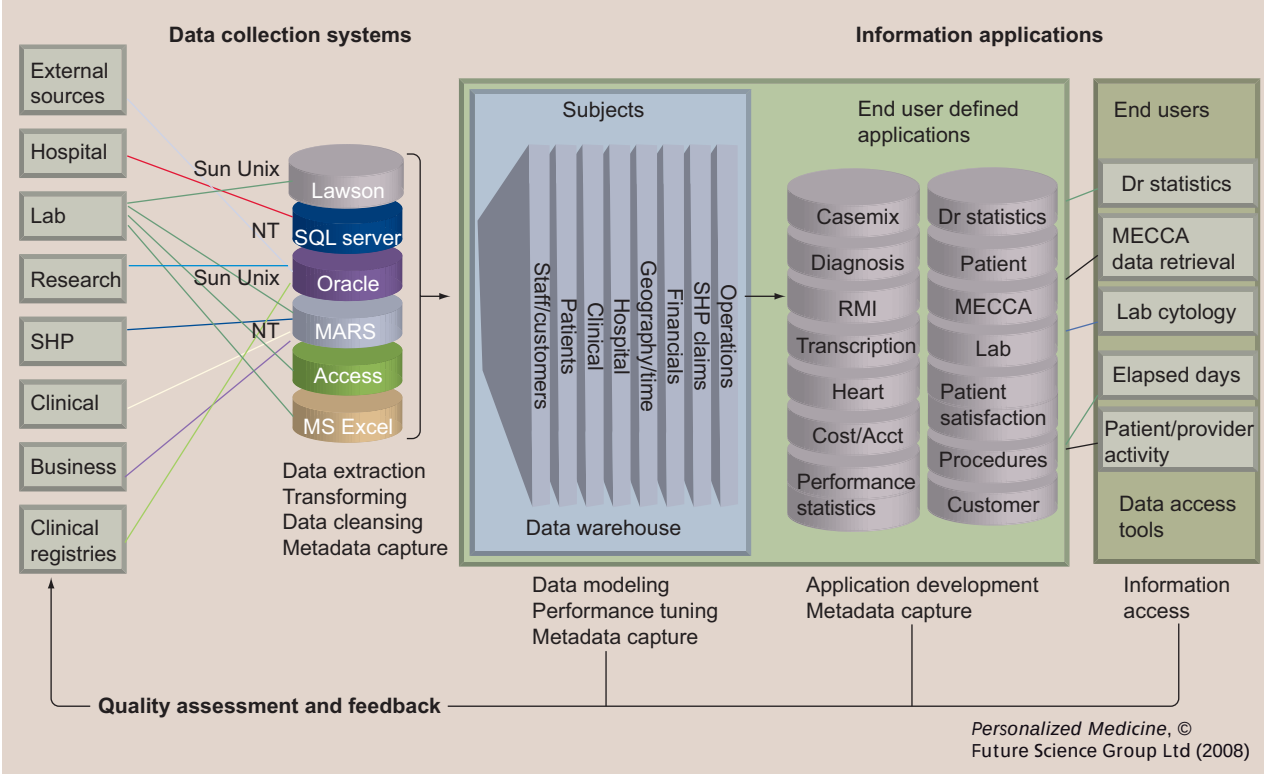
Personalized Medicine Research Project data represent a second data flow (shown in green in Figure 2). The movement of these data is similar to the clinical data flow, except that the data are collected as part of the PMRP. These data include information on a subject's environment, health habits, pedigree and family history. The PMRP data can either be packaged with the clinical data or packaged separately before sending to the encryption process.

When genetic samples are genotyped, the genetic data are deposited directly into the PMRD. A map between the genetic data sample identification number and the patient's encrypted identification number is kept so that all types of data can be combined for analysis. The PMRD has dramatically expanded in size since inception, and genotype data storage estimates are expected to exceed 10 terabytes over the next 2 years. The current database and data storage architecture will not support the estimated growth, thus the PMRP informatics team is evaluating options for future system growth.

Study infrastructure & additional data

Manolio and Collins of the National Human Genome Research Institute discussed the complexity of studying the interaction of genes and environment in health and disease [4]. Limited personal exposure information was collected on the original PMRP enrollment questionnaire (smoking and alcohol use primarily), in part to encourage a high participation rate, leaving the

Figure 1. The Marshfield Clinic data warehouse.



Redrawn with permission from Marshfield Clinic. © 2000.

possibility of recontact open through the written informed consent document. Given that dietary intake and physical activity are related to so many health outcomes, the strategic decision was made to retrospectively collect the data on enrolled subjects and to add the data collection prospectively with new enrollment. Marshfield Clinic is a partner with the University of Wisconsin in their Clinical and Translational Science Award and Institute for Clinical and Translational Research [102], and this collection of dietary intake and physical activity data was a major infrastructure project in years 1 and 2 of Clinical and Translational Science Award to increase the value of PMRP as a research resource for gene–environment studies. The Marshfield Clinic Institutional Review Board reviewed and approved the protocol.

Dietary intake

Consideration of several dietary assessment tools was undertaken. The gold standard, weighed food records, were deemed to be impractical due to the respondent burden and expense to quantify nutrient intake from the records. The 24-h recalls were also considered to be too burdensome to subjects and staff. Food frequency questionnaires (FFQs) are widely used to assess dietary intake in epidemiologic studies because they are more representative of usual intake and less expensive to implement than other methodologies, including weighed food records and 24-h dietary recalls, as they are usually self-administered [5,6]. The selected FFQ for the PMRP, the Diet History Questionnaire (DHQ) [103], was developed by researchers at the National Cancer Institute and has been shown to be superior to the commonly used Willett FFQ and similar to the Block FFQ for estimating absolute nutrient intakes [7–13]. The DHQ comprises 124 separate food items and asks about portion sizes for most foods. In addition, there are ten questions regarding nutrient supplement intake. The DHQ is printed and scanned by Optimum Solutions (CA, USA). After scanning, the data from the questionnaires will be stored in ASCII format and uploaded into the nutrient analysis software package. Diet*Calc software, available from the NIH, will be used for the nutrient analyses of the DHQ data [104].

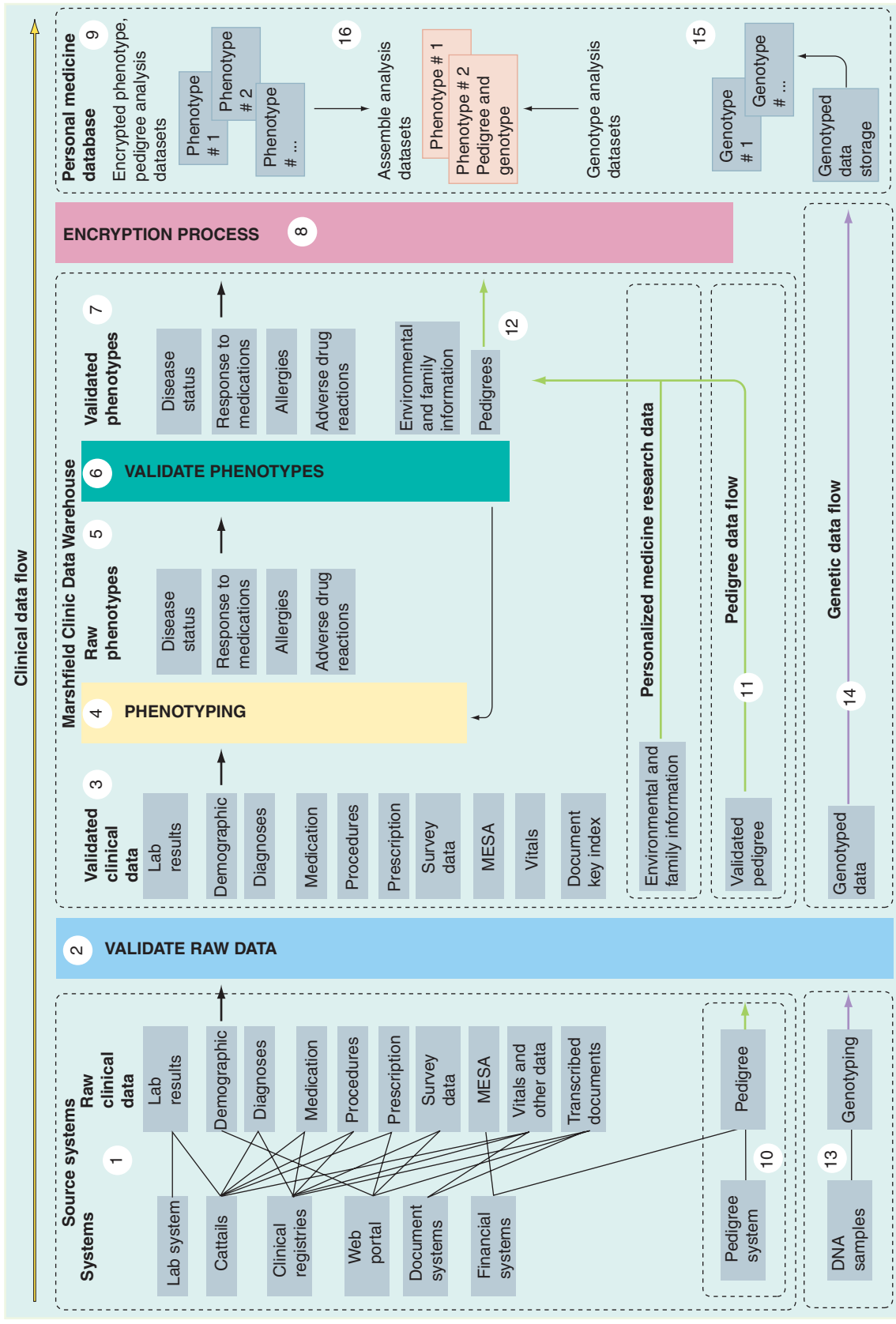


Figure 2. Process for moving data into the Personalized Medicine Research Database.

Redrawn with permission from Marshfield Clinic. © 2000.

Quantification of physical activity

As with measurement of dietary intake for epidemiologic studies, there are a number of different validated tools that have been used to measure physical activity in previous studies. The agreement between physical activity questionnaire and gold standard tends to be somewhat lower than for dietary intake, but is reasonable for ranking relative activity levels in groups. We have chosen to use a previously developed physical activity assessment tool to allow comparison with results from other study populations. Requirements of the selected tool for the current study included:

- Self-administered
- Previously validated
- Validated for use in a similar study population across a range of ages

The selected physical activity questionnaire, the Baecke questionnaire, is self-administered, validated for use in both men and women, and currently being used in a large, prospective study in the USA [14–17]. The questionnaire has been shown to have high reliability (coefficients ranging from 0.74 to 0.88) and accurate assessment of both high-intensity activity and light-intensity activity, such as walking. It comprises 16 questions and generates three indices of activity: a work index, a sport index and a leisure-time index.

The initial selection of PMRP subjects from which to obtain the DHQ and physical activity data included people with a BMI of 40 or over and two age- and sex-matched subjects with a BMI under 25 per obese subject. Two mailings and a follow-up telephone call led to a 65% response rate (n = 2103 completed questionnaires) in subjects who could be contacted. Summary DHQ data are presented in Table 4. Although average total daily caloric intake was not significantly different between the two groups, the data reveal significant differences in many of the other specific nutrient and food categories between obese and healthy weight subjects. Follow-up with the rest of the PMRP cohort to obtain these data is anticipated to be completed by the end of 2008.

The DHQ and physical activity questionnaires have been added prospectively for all new subjects being enrolled into PMRP. Subjects are given the questionnaires and asked to complete them and return them in stamped, self-addressed envelopes. With no further reminders, the response rate for prospective collection has been 74%, thus demonstrating the efficiency of collecting information at the time of enrollment.

Genetic association studies

The following sections outline the studies that are currently accessing the PMRP biobank under the categories of pharmacogenetics and genetic epidemiology. Although the PMRP was also designed to support studies of population genetics, to date there are no projects in this area. In addition to summarizing the study objectives, we have included infrastructure lessons learned through these studies. Results appear in separate peer-reviewed publications, as well as the PMRP study newsletters available on the PMRP website [101]. As suggested by the SAB in their early meetings, scientific discoveries using the PMRP have been enhanced through external collaborations, with a number of collaborations with members of the Pharmacogenetics Research Network [105].

Table 4. Comparison of daily dietary intake between healthy weight (BMI <25) and obese (BMI ≥40) subjects.

Nutrient	Median (5th percentile, 95th percentile)		p-value
	Healthy weight, n = 1249	***Obese, n = 529***	
Total energy (kcal)	1654 (815, 3749)	1711 (764, 3827)	0.23
Fat (%kcal)	32.0 (19.9, 42.9)	33.7 (21.8, 44.8)	<0.001
Saturated fat (%kcal)	10.7 (6.0, 16.7)	11.4 (7.1, 17.0)	<0.001
Protein (%kcal)	15.5 (10.6, 20.3)	16.7 (11.1, 21.4)	<0.001
Carbohydrate (%kcal)	52.2 (38.2, 67.5)	49.8 (37.0, 65.0)	<0.001
Calcium (mg)	844.7 (295.6, 2059.4)	819.8 (283.7, 2086.1)	0.35
Dietary fiber (g)	16.4 (6.7, 37.8)	16.1 (6.0, 38.8)	0.54
Iron (mg)	13.3 (5.9, 28.5)	13.2 (5.8, 29.1)	0.73
Vitamin A (µg)	759.5 (272.4, 1820.7)	764.5 (287.1, 1910.7)	0.01
Vegetable servings	3.2 (1.1, 8.9)	3.5 (1.0, 10.1)	0.01
Dairy servings	1.6 (0.3, 5.4)	1.5 (0.3, 5.3)	0.18
Grain servings	4.0 (1.6, 9.7)	4.1 (1.5, 10.5)	0.56

Table 5. Genetic epidemiology studies using the Personalized Medicine Research Project biobank, lessons learned and resulting publications.

Study	Purpose	External collaborations	Lessons learned	Publications
Glaucoma	To sequence a gene thought to be associated with glaucoma prevalence	Medical College of Wisconsin (WI, USA)	The scientific merit review process for projects requesting internal funding is too rigid for studies that do not require funding, and a streamlined process for PMRP studies is in development	None, as the project was only approved in 2008
AD	To identify candidate genes for AD and quantify gene–environment interactions with statin use, cigarette smoking and BMI	None	The original sampling frame (adults living in their own homes) precluded us from studying diseases for which people typically no longer live in their own homes. Subsequently, enrollment was opened to include nursing-home residents. Another lesson learned with this project, the first to genotype PMRP samples, was the need for more quality assurance as the samples are processed. One error in the sex markers was identified in this study of 450 subjects, and highlighted the need to develop a molecular fingerprint to identify the samples. The results of this project will be reported separately. Another lesson learned in this project was the amount of time necessary to characterize drug exposure from the mining of electronic notes for indication of drug use. Finally, we discovered that the smoking questions in the original PMRP questionnaire did not allow for the calculation of pack years because they did not include a question about how long subjects had smoked. Subsequently, postcards were mailed to the nearly 10,000 PMRP subjects who reported that they had ever smoked cigarettes to collect information about duration of smoking.	In preparation
FMS	To study a candidate gene for FMS and quantify gene–environment interaction with cigarette smoking, motor vehicle accidents and other trauma	None	Initial feasibility estimates overestimated the number of FMS cases in the PMRP cohort when applying strict diagnostic criteria. The project timeline was altered to allow for targeted recruitment of FMS subjects from zip codes other than the original 19 zip codes selected for PMRP enrollment	In preparation
Osteoporosis	To identify candidate genes for osteoporosis and to quantify potential gene–environment interactions with cigarette smoking and statin use	PGRN colleagues	Many women had not undergone BMD testing to confirm control status. An article was written for the twice-yearly PMRP newsletter describing the study and informing women of screening guidelines related to BMD testing in women aged 50 years and older	In preparation

AD: Alzheimer's Disease; BMD: Bone mineral density; CAD: Coronary artery disease; FMS: Fibromyalgia syndrome; GWAS: Genome-wide association studies; HDL: High-density lipoprotein; MI: Myocardial infarction; PMRP: Personalized Medicine Research Project.

Table 5. Genetic epidemiology studies using the Personalized Medicine Research Project biobank, lessons learned and resulting publications (cont.).

Study	Purpose	External collaborations	Lessons learned	Publications
CAD	To serve as a replication cohort for GWAS to identify markers associated with CAD	Medical College of Wisconsin	Changes in the reading of carotid ultrasonography over time necessitate manual review of charts to classify case–control status	None yet, as the study was only recently funded
Cataract and low HDL	To develop electronic algorithms for cataract and low HDL, to conduct community consultation activities related to data sharing, and to conduct a GWAS of cataract and low HDL	University of Wisconsin (WI, USA), Medical College of Wisconsin, Vanderbilt University (TN, USA)	The amount of disc space to store and manage data from GWAS is substantial. Although the PMRP consent form allows for data sharing, it needs to be more explicit about the data-sharing requirements of GWAS funded by the NIH	Community consultation manuscript under review
Dyslipidemia in severely obese subjects	To study candidate genes associated with diabetes and dyslipidemia in subjects with a BMI of 40+ and identify potential gene–environment interactions with diet and activity	Medical College of Wisconsin	A standard template for Material Transfer Agreements speeds the legal process to allow the shipment of samples to other laboratories for genotyping	None yet, as the project has only recently been initiated
Hypertensive heart disease	To identify candidate markers in the vitamin D pathway that increase risk for hypertension and hypertensive heart disease	University of Michigan (MI, USA)	Replication of findings from animal models is a sound strategy	In preparation
Myocardial infarction	To replicate two SNPs on 9p21 identified in GWAS for MI [21] in the context of known clinical covariates, including age, sex, smoking, physical activity, diabetes, BMI, family history of MI, and hypertension	Medical College of Wisconsin, University of Wisconsin	The first study to use all 20,000 PMRP samples, this study benefited from a master plating project that significantly reduced the time to create plates to ship off site for genotyping	None yet, as the project was only recently initiated

AD: Alzheimer's Disease; BMD: Bone mineral density; CAD: Coronary artery disease; FMS: Fibromyalgia syndrome; GWAS: Genome-wide association studies; HDL: High-density lipoprotein; MI: Myocardial infarction; PMRP: Personalized Medicine Research Project.

Table 6. Pharmacogenetic studies using the PMRP biobank, lessons learned and resulting publications.

Study	Purpose	External collaborations	Lessons learned	Publications
Topical β-blockers for glaucoma	To identify candidate markers that predict a 20% or greater drop in intraocular pressure	Medical College of Wisconsin (WI, USA)	Conservative definitions for windows of follow-up data create homogenous datasets, while decreasing sample size	[22–24]
Statin efficacy	To identify genetic markers of LDL response to atorvastatin	Pharmacogenetics Research Network colleagues (Children's Hospital Oakland Research Institute, CA, USA)	Allow additional time for MTAs that need to be reviewed and approved by multiple institutions. Master DNA plates dramatically decrease the amount of time necessary to identify and plate DNA samples for genotyping	[25,26]
Metformin for diabetes	To evaluate candidate markers for HbA1c response to metformin in Caucasians from PMRP and African–Americans from Kaiser Georgia (GA, USA)	Pharmacogenetics Research Network (University of California, San Francisco, CA, USA), Kaiser Georgia	As the PMRP is 98% Caucasian, collaboration with other groups is essential to allow replication of findings in other race/ethnic groups	
Warfarin	To identify predictors of stable warfarin dose and develop a dosing calculator to include all known modifiers of stable dose	Pharmacogenetics Research Network (University of Florida, FL, USA), International Warfarin Consortium	As recommended by the SAB, this was the first study to showcase the potential of the Marshfield biobank and to develop many external collaborations, including one to identify new compounds for clinical testing. These studies have led to several peer-reviewed publications, the discovery of a new genetic association with warfarin stable therapeutic dose, funding from the Agency for Healthcare Research and Quality to prospectively evaluate the efficacy of gene-based prescribing for warfarin, and the development of the IWPC, an international consortium aggregating 6000 patient's data to study warfarin pharmacogenetics. The data from the IWPC will be placed in the public domain through the PharmGKB website [109] in late 2008	[27–31]
Tamoxifen in breast cancer	To identify makers for response and adverse reactions (deep-vein thrombosis) to tamoxifen for breast cancer	Pharmacogenetics Research Network (Indiana University, IN, USA), Medical College of Wisconsin	A standard template MTA speeds the legal process	Under review
Statin myotoxicity and rhabdomyolysis	To identify genetic markers that predict the most common toxicity for one of the most commonly prescribed drugs in the world	Pharmacogenetics Research Network (Children's Hospital Oakland Research Institute), Medical College of Wisconsin	Statin myotoxicity is very difficult to phenotype using medical records because the documentation of pain symptoms is not standard	[32,33]

IWPC: International Warfarin Pharmacogenetic Consortium; LDL: Low-density lipoprotein; MTA: Material Transfer Agreements; PMRP: Personalized Medicine Research Project; SAB: Scientific Advisory Board.

Genetic epidemiology

In addition to standard case–control studies of disease onset, the PMRP is ideally suited to the study of risk factors for disease progression, because over 20 years of medical history information is available for approximately 75% of the cohort [18]. Table 5 summarizes the genetic epidemiology studies using the PMRP database and the many lessons learned.

Pharmacogenetic studies

Prior to the implementation of electronic prescribing and a paperless environment in 2007, mention of medication use in the text was captured through natural language processing and is available back to 1993. These resources have allowed the initiation of many pharmacogenetic studies, several of which are in collaboration with colleagues in the Pharmacogenetics Research Network [105]. As with genetic epidemiology, many lessons have been learned to increase efficiency and accuracy (Table 6).

Conclusion

In summary, much has been learned since the PMRP was launched in 2002, and many discovery projects are underway. With each new study undertaken, it is important to revisit procedures to be sure that they keep pace with genetic and statistical tools that are being developed and refined. We have found that it is far more efficient to collect personal exposure and environmental data at the time that people are enrolled into the biobank and to conduct quality assurance on the samples as they are received. Collaboration has greatly enhanced the science, and multidisciplinary teams are essential to address the complex issues involved with studying complex diseases. As new biobanks are being initiated, the sharing of methods and standards will help to facilitate sharing of data for comparison and replication.

Future perspective

Biobanking

Banking of biological materials for research and clinical purposes will likely continue to increase and will decrease the time between discovery and clinical translation. Much can be learned by sharing the success and failures of these endeavors. The PMRP has been committed to sharing data, and lessons learned, for the wider scientific community.

The current National Human Genome Research Institute-funded electronic medical records and genomics (eMERGE) network [106], of which PMRP is a member, is charged with developing standards to allow for the efficient sharing of data from across the network. The eMERGE network is a national consortium formed to develop, disseminate and apply approaches to research that combine DNA biorepositories with electronic medical record systems for large-scale, high-throughput genetic research. Through researching new methodologies and disseminating the information, this consortium will improve the use of biorepositories for genomic research. The five sites funded included Marshfield Clinic, Northwestern University (IL, USA), Vanderbilt University (TN, USA), Mayo Clinic (MN, USA) and Group Health Cooperative (WA, USA).

Organizations such as the Public Population Project in Genomics (P3G) [107] are also seeking to develop best practice for biobanking. P3G is a not-for-profit international consortium to promote collaboration between researchers in the field of population genomics. It was launched to provide the international population genomics community with the resources, tools and know-how to facilitate data management for improved methods of knowledge transfer and sharing. Its main objective is the creation of an open, public and accessible knowledge database. The motto is 'transparency and collaboration'. All groups benefit from the sharing of methodologies and data.

The challenge for all of these groups is to identify infrastructure funding and to ensure timely translation of research results.

Translation into personalized healthcare

Small- and large-scale genetic association studies are moving expert biological knowledge from the bench to the bedside at an unprecedented rate. The Personalized Health Care Initiative from the Department of Health and Human Services seeks to "improve the safety, quality and effectiveness of healthcare for every patient in the US by using genomics … to enable medicine to be tailored to each person's needs" [108]. The substrate of functional genomics is expanding geometrically, and efficient bioinformatics resources are being developed feverishly in an effort to help scientists and clinicians leverage this knowledge to improve healthcare. The potential seems unlimited. Genetic risk determinants can be merged with clinical data to predict patient risk, before the onset of disease. This should allow focused early intervention for patients at highest risk. Genetic markers can also be used for risk stratification once a particular condition has developed. This should

optimize secondary prevention and reduce the burden of comorbidity being placed upon our healthcare infrastructure as the population continues to age. Genetic markers will also be highly useful in directing therapy, in the context of pharmacologic as well as nonpharmacologic intervention. For drugs with a narrow therapeutic index, this has already become a clinical reality. Many antineoplastic agents are prescribed and/or dosed based upon genotype in an effort to reduce the risk of adverse drug reactions, and evidence is mounting that this approach will be beneficial for other drugs, such as anticoagulants. As technology advances (e.g., genome-wide association studies), the genetic architecture underlying drug response will almost certainly become more fully characterized for a multitude of therapeutic agents, making this prospective gene-based approach to risk reduction and treatment prescription advantageous for all drugs, including those with a relatively wide therapeutic index [19,20]. Proteomics and metabolomics will further the usefulness of genomics discovery for personalized healthcare. The vast amount of information necessary for discovery and decision support requires a robust informatics infrastructure. A well-designed and -executed biobank is merely the first step in the path from basic discoveries to personalized healthcare.

Acknowledgements

The authors acknowledge Dr Laura Coleman and Bickol Mukesh for their contributions to the analysis of the dietary intake data.

Financial & competing interests disclosure

The research was funded, in part, by grant number 1 D1A RH00025-01 from the Office of Rural Health Policy, Health Resources and Services Administration, from the Department of Commerce (WI, USA) and grant number 1UL1RR025011 from the Clinical and Translational Science Award (CTSA) program of the National Center for Research Resource National Institute of Health. The members of the PMRP Scientific Advisory Board include: David Altshuler, MD, PhD, Massachusetts Institute for Technology (MA, USA), Chair; David Flockhart, MD, PhD, Indiana University (IN, USA); Stephen Liggett, MD, University of Maryland (MD, USA); Gabor Marth, DSc, Boston College (MA, USA); Jurg Ott, PhD, Rockefeller University (NY, USA); Lenna Peltonen, MD, PhD, Biomedicum Helsinki (Helsinki, Finland); Wendell Weber, MD, PhD, University of Michigan (MI, USA). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Scientific Advisory Board

- The Scientific Advisory Board was very useful in advising on design and strategic issues for growth and development, and it would have been useful to have their advice on issues that developed as the biobank grew, such as new technologies. Infrastructure funding to support the continued engagement of the Scientific Advisory Board has been difficult to identify. Institutional support from the Marshfield Clinic on an annual basis must be prioritized, and has been used to continue recruitment and to develop the bioinformatics tools necessary to access the biobank.

Oversight Committee

- It is neccessary to develop the process and procedures to access the biobank and make forms available online. The Oversight Committee should have representation from content experts in the types of samples available, as well as the informatics support crucial to the studies.

Sample tracking

- It is neccessary to outline the computer hardware and software needs to manage an ever-growing genotypic and phenotypic dataset that allows for easy searchability. The advantages and disadvantages of developing the necessary software in-house or purchasing a system and having it customized as needed must be weighed up.

Study infrastructure & additional data

- The collection of personal/environmental exposure data at the time of study enrollment should be considered. Fewer resources are required to collect the information at the time of enrollment, and the participation will be higher then. However, the amount of data collected should be weighed against the potential for a lower participation rate initially because of respondent burden.

Genetic association studies

- Many lessons have been learned in the conduct of the genetic discovery studies, as outlined in Tables 5 & 6. Constant review of procedures to keep pace with changes in technology is essential.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Kaiser J: Population databases boom, from Iceland to the U.S. *Science* 298, 1158–1161 (2002).
2. Collins FS, Green ED, Guttmacher AE, Guyer MS: A vision for the future of genetics research. A blueprint for the genomic era. *Nature* 422, 835–847 (2003).

•• Important paper from the scientific leadership at the National Human Genome Research Institute, outlining a broad strategy for genetics research after the completion of the Human Genome Project.

3. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD: Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Med.* 2, 49–79 (2005).
4. Manolio TA, Collins FS: Genes, environment, health, and disease: facing up to complexity. *Hum. Hered.* 63, 63–66 (2007).
5. Willett W: *Nutritional Epidemiology*. Oxford University Press, NY, USA (1990).
6. Margetts BM, Nelson M (Eds): *Design Concepts in Nutritional Epidemiology*. Oxford University Press, Oxford, UK (1991).
7. Subar AF, Thompson FE, Kipnis V *et al.*: Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. *Am. J. Epidemiol.* 154, 1089–1099 (2001).
8. Millen AE, Midthune D, Thompson FE, Kipnis V, Subar AF: The National Cancer Institute diet history questionnaire: validation of pyramid food servings. *Am. J. Epidemiol.* 163, 279–288 (2005).
9. Flood A, Subar AF, Hulll SG, Zimmerman TP, Jenkins DJA, Schatzkin A: Methodology for adding glycemic load values to the National Cancer Institute diet history questionnaire database. *J. Am. Diet. Assoc.* 106, 393–402 (2006).
10. Subar AF, Midthune D, Kulldorff M *et al.*: Evaluation of alternative approaches to assign nutrient values to food groups in food frequency questionnaires. *Am. J. Epidemiol.* 152, 279–286 (2000).
11. Thompson FE, Subar AF, Brown CC *et al.*: Cognitive research enhances accuracy of food frequency questionnaire reports: results of an experimental validation study. *J. Am. Diet. Assoc.* 102, 212–225 (2002).
12. Subar AF, Ziegler RG, Thompson FE *et al.*: Is shorter always better? Relative importance of questionnaire length and cognitive ease on response rates and data quality for two dietary questionnaires. *Am. J. Epidemiol.* 153, 404–409 (2001).
13. Subar AF, Thompson FE, Smith AF *et al.*: Improving food frequency questionnaires: a qualitative approach using cognitive interviewing. *J. Am. Diet. Assoc.* 95, 781–788 (1995).
14. Baecke JAH, Burema J, Frijters JER: A short questionnaire for the measurement of habitual physical activity in epidemiological studies. *Am. J. Clin. Nutr.* 36, 936–942 (1982).
15. Richardson MT, Ainsworth BE, Wu H-C, Jacobs DR, Leon AS: Ability of the Atherosclerosis Risk in Communities (ARIC)/Baecke questionnaire to assess leisure-time activity. *Int. J. Epidemiol.* 24, 685–693 (1995).
16. Pols MA, Peeters PHM, Bueno-de-Mesquita HB *et al.*: Validity and repeatability of a modified Baecke questionnaire on physical activity. *Int. J. Epidemiol.* 24, 381–387 (1995).
17. Tehard B, Saris WHM, Astrup A *et al.*: Comparison of two physical activity questionnaires in obese subjects: the NUGENOB study. *Med. Sci. Sport Exerc.* 37, 1535–1541 (2005).
18. McCarty CA, Mukesh BN, Giampietro PF, Wilke RA: Healthy People 2010 disease prevalence in the Marshfield Clinic Personalized Medicine Research Project: opportunities for public health genomic research. *Personalized Med.* 4, 183–190 (2007).
19. Wilke RA, Reif DG, Moore JH: Combinatorial pharmacogenetics. *Nat. Rev. Drug Discov.* 4, 911–918 (2005).
20. Wilke RA, Mareedu RK, Moore JH: The pathway less traveled – moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies. *Curr. Pharmacogenom.Pers. Med.* (2008) (In Press).
21. Helgadottir A, Thorleifsson G, Magnusson KP *et al.*: The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nat. Genet.* 40, 217–224 (2008).
22. McCarty CA, Mukesh BN, Kitchner TE *et al.*: Intraocular pressure response to medication in a clinical practice-based cohort – the Marshfield Clinic Personalized Medicine Research Project (PMRP). *J. Glaucoma.* (2007) (In Press).
23. McCarty CA, Burmester JK, Mukesh BN, Patchett R, Wilke RA: Intraocular pressure response to topical β-blockers is associated with an *ADRB2* SNP. *Arch. Ophthalmol.* 126, 959–963 (2008).
24. Sidjanin DJ, McCarty CA, Patchett R, Smith EM, Wilke RA: Pharmacogenetics of ophthalmic β-blockers. *Personalized Med.* 5, 377–385 (2008).
25. Peissig P, Sirohi E, Berg RL *et al.*: Construction of atorvastatin dose–response relationships using data from a large population-based DNA biobank. *Basic Clin. Pharmacol. Toxicol.* 100, 286–288 (2007).

• Describes our approach to electronically characterize medication exposure for pharmacogenetics studies.

26. Wilke RA, Berg RL, Linneman JG, Zhao C-F, McCarty CA, Krauss RM: Characterization of LDL-cholesterol lowering efficacy for atorvastatin in a population-based DNA biorepository. *Basic Clin. Pharmacol. Toxicol.* (2008) (In Press).
27. Hillman MA, Wilke RA, Caldwell MD, Berg R, Glurich I, Burmester JK: Relative impact of covariates in prescribing warfarin according to *CYP2C9*-based genotype. *Pharmacogenetics* 14, 539–547 (2004).
28. Wilke RA, Berg RL, Vidaillet H, Caldwell MD, Burmester JK, Hillman MA: Impact of age, *CYP2C9* genotype, and concomitant medication on the rate of rise for prothrombin time, during the first 30 days of warfarin therapy. *Clin. Med. Res.* 3, 207–213 (2005).
29. Hillman MA, Wilke RA, Yale S *et al.*: A prospective, randomized pilot trial of model-based warfarin dose initiation using *CYP2C9* genotype and clinical data. *Clin. Med. Res.* 3, 137–145 (2005).

• First paper describing the results of a prospective study to evaluate the results of gene-based prescribing for warfarin.

30. Caldwell MD, Berg RL, Zhang KQ *et al.*: Evaluation of genetic factors for warfarin dose prediction. *Clin. Med. Res.* 5, 8–16 (2007).
31. Caldwell MD, Awad T, Johnson JA *et al.*: *CYP4F2* genetic variant alters required warfarin dose. *Blood* 111, 4106–4112 (2008).
32. Wilke RA, Moore JH, Burmester JK: Relative impact of *CYP3A* genotype and concomitant medication on the severity of atorvastatin-induced muscle damage. *Pharmacogenet. Genomics* 15, 415–421 (2005).
33. Wilke RA, Lin D, Roden DM *et al.*: Identifying genetic risk factors for serious adverse drug reactions – current progress and challenges. *Nat. Rev. Drug Discov.* 6, 904–916 (2007).

Websites

101. Marshfield Clinic Research Foundation Personalized Medicine Research Project www.marshfieldclinic.org/pmrp
102. University of Wisconsin Institute for Clinical and Translational Research www.ictr.wisc.edu
103. US NIH, National Cancer Institute: Risk factor monitoring and methods: diet history questionnaire (2007) http://riskfactor.cancer.gov/DHQ
104. US NIH, National Cancer Institute: Risk factor monitoring and methods: diet history questionnaire: Diet*Calc Software (2007) http://riskfactor.cancer.gov/DHQ/dietcalc
105. US NIH, National Institute of General Medical Sciences: Pharmacogenetics research network (2008) www.nigms.nih.gov/Initiatives/PGRN
106. The eMERGE Network – Electronic Medical Records and Genomics www.gwas.org

•• Public website for the eMERGE network, a consortium of five sites funded by the National Human Genome Research Institute to develop, disseminate and apply approaches to research that combine DNA biorepositories with electronic medical record systems for large-scale, high-throughput genetic research.

107. Pharmacogenetics Research Network www.nigms.nih.gov/Initiatives/PGRN

•• Includes information about the Pharmacogenetics Research Network, which is funded by the NIH. The goal of the network is to discover how the genes that vary among individuals affect drug safety and efficacy.

108. HHS.gov: Personalised Health Care www.dhhs.gov/myhealthcare

•• Released in September 2007, this document outlines opportunities, challenges, pathways and resources to make personalized healthcare a reality through expansion of the science base, health information technology, intervention development and review, and integration into clinical practice.

109. PharmGKB: The Pharmacogenetics and Pharmacogenomics Knowledge Base www.pharmgkb.org