

Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine

DM Roden¹⁻³, JM Pulley⁴, MA Basford^{1,4}, GR Bernard^{2,4}, EW Clayton^{5,6}, JR Balsler^{3,4} and DR Masys⁷

Our objective was to develop a DNA biobank linked to phenotypic data derived from an electronic medical record (EMR) system. An “opt-out” model was implemented after significant review and revision. The plan included (i) development and maintenance of a de-identified mirror image of the EMR, namely, the “synthetic derivative” (SD) and (ii) DNA extracted from discarded blood samples and linked to the SD. Surveys of patients indicated general acceptance of the concept, with only a minority (~5%) opposing it. As a result, mechanisms to facilitate opt-out included publicity and revision of a standard “consent to treatment” form. Algorithms for sample handling and procedures for de-identification were developed and validated in order to ensure acceptable error rates (<0.3 and <0.1%, respectively). The rate of sample accrual is 700–900 samples/week. The advantages of this approach are the rate of sample acquisition and the diversity of phenotypes based on EMRs.

The notion that common human traits, such as susceptibility to disease, include a genetic component is ingrained in human history and literature. The idea that genetic factors might similarly modulate response to exogenous substances, including drugs and food, was proposed in the early twentieth century by the great English physiologist Garrod in his studies of “inborn errors of metabolism.”¹ This notion preceded by 50 years the first descriptions of striking variability in response to drug therapy caused by single gene variants,²⁻⁵ and the second half of the twentieth century has seen a dramatic increase in our understanding of the way in which such variants cause human disease⁶⁻¹⁰ and modulate responses to drugs.^{11,12}

These ideas are now evolving from mere intellectual curiosities applicable to only a few clinical settings to part of the fabric of modern medical practice. The rapid evolution of technologies capable of acquiring and analyzing high-dimensionality data (genomes, proteomes, and images) now holds the potential for dramatically extending this line of reasoning to make it the centerpiece of a widely heralded era of personalized medicine^{13,14} in which not only disease treatment but also preventive therapies will be applied in a personalized fashion.

Identifying robust genotype–phenotype relationships is a key challenge in the process of making this vision a reality. This requires very large sample sets for discovery and validation.⁶⁻¹⁰ This challenge, in turn, is the compelling rationale for the

establishment of interoperable biobanks across the world.¹⁵⁻¹⁷ The extent to which such resources represent the populations from which they are drawn is uncertain; however, coupling these biobanks to electronic medical record (EMR) systems has the potential to enable investigators in the field of genomics to search, record, and analyze phenotypic information pertaining to large numbers of patients in a “real world” context.^{18,19} We describe here a method to generate an “opt-out” system based on the use of blood samples collected for clinical purposes and subsequently discarded. The approach is predicated on extensive ethics and community input and unprecedented systematic de-identification of an entire EMR.

RESULTS

Regulatory and ethics review

The Vanderbilt Institutional Review Board (IRB) reviewed the initial project plan and agreed that it met the criteria to be designated as “nonhuman subjects” research. However, given the anticipated scale of the project and its potential impact on the community, the IRB advised additional safeguards. These included ongoing institutional and IRB oversight; evaluation by the Medical Center’s Ethics committee; and establishment of Ethics, Scientific, and Community Advisory Boards. The Medical Center Ethics Committee review generated recommendations (**Table 1**) for modifications and further evaluation of the program before its

¹Office of Personalized Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA; ²Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA; ³Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA; ⁴Office of Research, Vanderbilt University School of Medicine, Nashville, Tennessee, USA; ⁵Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA; ⁶School of Law, Vanderbilt University, Nashville, Tennessee, USA; ⁷Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. Correspondence: DM Roden (dan.roden@vanderbilt.edu)

Received 11 December 2007; accepted 27 March 2008; advance online publication 21 May 2008. doi:10.1038/cpt.2008.89

Table 1 Key steps in implementations of the Vanderbilt DNA databank

Develop sample handling programs
Develop the synthetic derivative
Implement and validate de-identification methodology
Convene ethics review and act on recommendations:
Ongoing IRB oversight
All patients should receive information about the project
Patients should have the right to refuse to participate
Develop a comprehensive education plan
Ongoing external evaluation
Assess the content and functionality of the database before enrolling participants
Establish Ethics Advisory Board and Scientific Advisory Board
Involve clinical operations and legal staff

IRB, Institutional Review Board.

launch, and these are described later. All of these reviews were transparent and available to other advisory bodies. Specific Office for Human Research Protections opinions were sought during the setting up of the resource and immediately before the launch in its final form so as to ensure that the resource was consistent with federal regulations for exempt research.

Surveys of patients: In response to the mail survey sent out to patients,²⁰ 90% of the respondents were comfortable with the idea of anonymized genetic information being used for research, and a smaller group (~5%) was opposed to it. This finding suggested a need for notification. In-person interviews with patients post-phlebotomy suggested that only a minority (32%) of the patients who had seen the poster could recall it; however, after they had been read a brief statement about the program, most of them (93%) reported that they were comfortable with the idea of the DNA databank concept.²¹ The findings from these studies underlined the need to communicate with patients about the program, as well as to allow an easy and permanent opt-out mechanism.

Changes in the standard “consent to treatment” form: On the basis of the data from the survey of patients and the interpretation of these data by the advisory boards, the standard “consent to treatment” form was modified to include a statement in bold lettering describing the DNA databank concept; also, a box to check for opting out was placed directly above the signature lines. In addition, a new institutional policy was introduced so that this form is now signed not only at each hospital admission but also by each outpatient on an annual basis. Patients can also opt out by calling a specific phone number. The sample handling program was modified to include only subjects with a signed “consent to treatment” document that did not include a tick in the box. The opt-out rate is ~2.5% of all patients who sign the form.

Interactions with the community

A Community Advisory Board was established to ensure community involvement and input into the design and function

of the repository operations, with the goal of evaluating and ultimately supporting acceptance among broader medical as well as lay audiences. The Board has 12 members representing diverse sections of the community, who play an active role in the community through employment, parenting activities, church groups, civic groups, educational activities, or extracurricular activities. A familiarity with technology, science, or genetics is not expected. Their specific responsibilities are to evaluate the ethical conduct of the repository’s operations in the context of the security and privacy measures that are in place; to act as the voice of the community on any issues relating to the use of genetic information for research; and to identify core ethical or social dilemmas and propose concrete and practical measures toward resolving them.

Tests of sample handling and interface with the synthetic derivative

The gender-matching test identified 8/384 gender mismatches between the electronic record and the genotyping results. On review, these mismatches were found to include three patients who had received hypertransfusion in the recent past, three with a history of bone marrow/stem cell transplant, and one with no information in the associated medical record (this can occur, for example, with “research only” samples sent to the medical center). The residual error rate was therefore 1/384 (0.26%). On the basis of this analysis, sample handling procedures were altered so as to flag subjects who had undergone hypertransfusion or bone marrow transplant for the information of investigators requesting those samples/data.

In an experiment dealing with sickle cell/macular degeneration, there were 296/300 (98.7%) concordant results and 4 mismatches; of the mismatches, 1 patient had received blood transfusions in the recent past and 2 had very limited data in the synthetic derivative (SD) (likely “research only” patients), leaving a residual error rate of 0.3%.

The results with gender typing and allele identification in sickle cell disease and age-related macular degeneration suggest that the resource should, on a large scale, be evaluated for the analysis of complex genotype–phenotype relationships.

The SD

The developed de-identification algorithm removed 5,378 of the 5,472 identifiers, with an error rate for complete Health Insurance Portability and Accountability Act (HIPAA) identifiers of <0.1%. The aggregate error rate—which includes any potential error, including non-HIPAA items, partial items, and items that are not inherently related to identity—was 1.7% (95% confidence interval: 1.4–2.1%). Examples of over- and underscrubbing are shown in [Figure 1](#). The names of patients would typically be the most inherently identifying part of the record. The process by which names are scrubbed involves a complementary utilization of census-derived name dictionaries as well as a matching function based on the name information contained in the header files of the original, unscrubbed EMRs. In the initial analysis, the error rate for all names contained in the record was 3% (95% confidence interval: 2.1–3.8%). However,

all the unscrubbed names were those of health-care providers, and no names of patients were unscrubbed. After that result, a copy of the institutional employer dictionary was added to the name-scrub process. Given the very large number of records, it is possible that a patient's name will be undermarked within a given scrubbed record; however, such undermarking would be either the first name or the last name and would be very unlikely to include both names. Because de-identification algorithms are not (and never can be) perfect, the resource is considered to be a limited dataset as defined by the HIPAA privacy rule, and its use is therefore governed by the terms of a data use agreement signed by investigators when they query the SD, as described

later in this article. In addition, when individual undermarking errors are uncovered and reported, they can be permanently fixed. This general concept of re-identifiability is generic to the routine use of de-identified health-care information in research and is not unique to the Vanderbilt DNA databank.

Because the de-identification procedures can discover and suppress identifiers with a high degree of accuracy, the IRB judged them to be consistent with an OHRF "nonhuman subjects" designation. The algorithms were applied to the EMR to create the SD, and information accrued in the EMR is added to the SD once a week. Figure 2 illustrates the scrubbing applied to an SD record.

Prescrub	After scrub	Error type
Undermarking		
Rx for Lortab 10, #60 w/ one refill 12/8/4	Rx for Lortab 10, #60 w/ one refill 12/8/4	Misformatted date not recognized
The number of the ventilator is 98141, patient being monitored with oximetry	The number of the ventilator is 98141, patient being monitored with oximetry	Device ID not recognized
Overmarking		
GI: soft, ND, normal bowel sounds, non tender, no hepatomegaly, no splenomegaly	GI: **PLACE, ND, normal bowel sounds, non tender, no hepatomegaly, no splenomegaly	ND misidentified as state abbreviation
With iron, 40 g protein daily, and 1,500–2,000 calories daily	With iron, 40 g protein daily, and **ID-NUM calories daily	Number misidentified as ID
An attending cardiologist was present throughout the diagnostic study	An attending **NAME [SSS] was present throughout the diagnostic study	Descriptor misrecognized as name

Figure 1 Examples of under- and overmarking. The original text is shown on the left and the result of the scrubbing process is shown in the middle. The target text and the result of scrubbing are highlighted in red. (Lortab; Mikart, Atlanta, GA.)

Status

Sample collection was launched in February 2007 after a 3-year planning process to develop and refine the project. The accrual rate has been 700–900 samples/week, adding up to 33,463 samples as of April 2008; by the end of 2010, the resource will include >130,000 samples. The de-identified SD records of the first 16,102 samples that were accepted contained a mean of 14 ± 18 ICD-9 codes. The most common diagnoses were hypertension (15.7%), type II diabetes (11.8%), hyperlipidemia (11.5%), coronary artery disease (7.8%), and anemia (5.9%). The records have a mean of 6.5 ± 6.2 years of history, with just over half the records (55%) containing documentation of an inpatient stay. Most (88%) records have at least one medication indicated, with an average of 8.0 ± 6.8 medications per record. In addition, 69% have one or more procedure codes. The vast majority have at least one lab report, consistent with the fact that the samples were collected from leftover blood from pathology.

A first-generation web-based tool has been developed that can determine the number of cases with specific investigator-defined

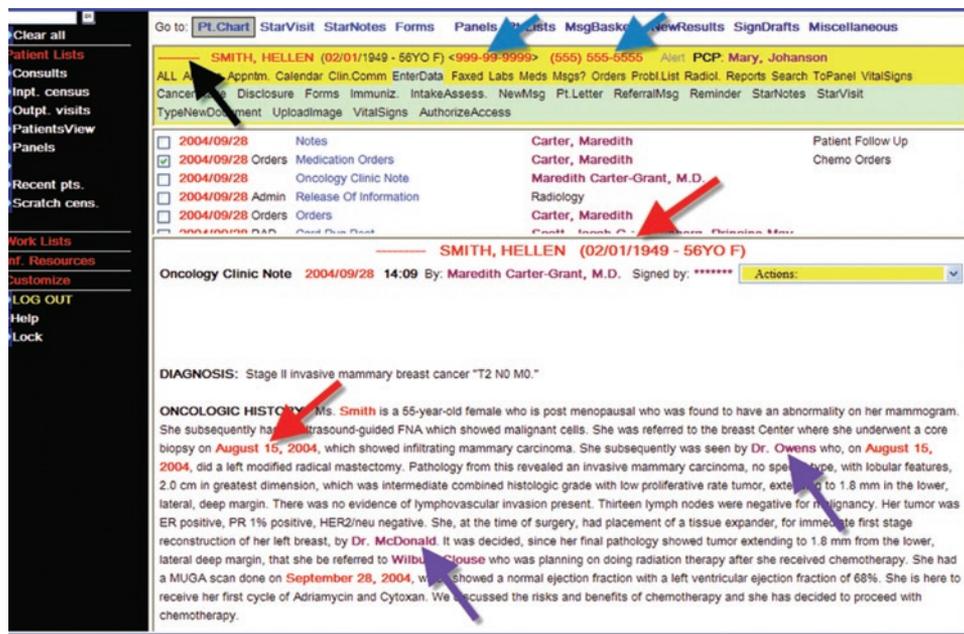


Figure 2 A descriptive example of a record in the synthetic derivative (SD) described in the text. The arrows indicate examples of scrubbing: the medical record number has been removed (black), the social security and phone numbers have been masked (blue), names have been changed (purple), and dates have been shifted (red) as described in Methods.

attributes that are present among the cases in the SD, and it can be accessed by any individual with a Vanderbilt web access password. A computer screen view of the first-generation query tool that examines cases identified by these queries is shown in [Figure 3](#).

Proposals for research projects that intend to use the DNA databank are IRB-reviewed and the potential users are required to sign the data use agreement, including an acknowledgement that genetic information comes with potential pathways for re-identification. Accordingly, users agree not to disclose data and to implement safeguards to prevent disclosure of the data. Unauthorized use, including any attempt to re-identify the information contained in the dataset or to perform unapproved research even on an approved dataset, carries administrative and institutional penalties. SD searches that return very small sample sets (currently ≤ 5) will not be released. The data use agreement further stipulates that all genotypic information generated will be re-deposited in the SD, thereby facilitating studies of genotype–phenotype relationships as the resource becomes increasingly well populated with genomic data.

DISCUSSION

A major focus in contemporary genome science has been the analysis of common phenotypes in large populations in order to determine alleles associated with risk. This effort, propelled by the rapid emergence of tools for genome-wide association, has been remarkably successful but requires very large study and replication sets.^{6–10} The obvious functions that can be fulfilled by the diagnosis-agnostic resource we describe include both hypothesis generation and replication.

In addition, the emergence of sets of many-times-replicated robust genotype–phenotype relationships poses the new challenge of establishing how this information can be integrated into routine health care. In this context, a second rationale for the resource we describe—that links a rich EMR to genomic information—is the development of tools and new knowledge in diverse areas such as biomedical informatics, evidence-based personalized medicine, point-of-care information delivery, ethics, and health-care economics that can help in bringing about such an integration.

Advantages of the opt-out approach. Two designs were initially considered, a conventional “opt-in” (consent) model, and the “opt-out” model that was ultimately adopted. An absolute requirement for the opt-out model is an EMR coupled to extensive informatics resources that allow the requisite de-identification. This was enabled at Vanderbilt by a well-established and locally developed EMR specifically designed for clinical care as well as clinical research and data mining.^{22–27}

The generation of prospectively ascertained study populations by the opt-in approach is resource-intensive and therefore often focuses on specific diseases or therapies. In contrast, the opt-out approach can generate very large datasets that include a rich diversity of phenotypes. Further, surveys indicate that the opt-in model may exclude large segments of the population,^{28–30} whereas, in the opt-out model we describe here, the resource has the potential to be more broad-based and is not limited by prospectively designed research questions. Indeed, a default opt-out recruitment approach has been advocated for all “low-risk” studies.³¹

Research involving DNA generates uneasiness among some sections of both science and lay communities because of the perceived potential for re-identification and the attendant potential for the unintended use of such identification (e.g., genetic discrimination or coercion by law courts to reveal DNA samples or data).^{32–34} An opt-in model allows participants to agree to the use of identified information for research; however, data containing genetic information housed within a de-identified resource has far less potential for unintended discriminatory use than an identified resource, even in the event of unlawful release. We therefore believe that there are specific advantages (despite the inconvenience and the cost, which are not inconsiderable in either model) to the approach we describe. We also recognize that the ability to utilize this resource would further safeguard the confidentiality of patients’ data by replacing fully identifiable records with IRB-approved protocols under a waiver-of-consent authorization.

There are other important challenges (e.g., technological, informatic, biostatistical) that must be addressed in order to achieve the goal of personalized medicine, and the resource

DNA Databank and Synthetic Derivative			
Search Criteria <input type="button" value="Change"/> <input type="button" value="Refine"/> <input type="button" value="Save"/>			
Diag Keys: *A*5b* and Hypertension	No: *Diabetes Mellitus*	Age: 25-50	DNA sample: Available
Med Keys: Ciprofloxacin and *Prednisone Oral*	No:	Gender: Male	Smoking: Current and former smoking
ICD-9 Codes: 282.61 and 719.*	No: 389.11	Ethnicity: AA	Records Dated: Last 5 years
CPT Codes: 92557	No: 99214	Marital Status: single	
Total result returned 137 <input type="button" value="Change"/> <input type="button" value="Refine"/> <input type="button" value="Save"/> <input type="button" value="Export"/>			
Male, 27, Single, AA, former lighter smoking, 282.6(ICD-9), 92557 (CPT), DNA2765092451			Select
2005/02/14 he has numerous health problems related to his morbid obesity including insulin-dependent diabetes, hypertension, among others. He presents for elective bypass. The more risky nature of repeat gastri			<input type="checkbox"/>
2002/11/04 Ciprofloxacin Oral Tablet 500 mg 1 tablet by mouth every 12 hours for ten days			<input type="checkbox"/>
Male, 38, Single, AA, current heavy smoking, 719.3(ICD-9), 92557 (CPT), DNA5870668975			Select
2000/05/05 Vision Impaired, Risk Alerts: Cardiac Ischemia: Moderate, Aspiration: High, Difficult Airway: High, Hypertension: Moderate Evaluator: **NAME[XXX, WWW] Evaluation Date: **DATE[May 23, 2005] Reviewer Co			<input type="checkbox"/>

Figure 3 Synthetic derivative interrogation tool. Search criteria are entered in the blue box, and entries and potential records are returned, with the “keywords in context” shown below. The user then has the option of including the record in the sample set to be analyzed. (Ciprofloxacin; Bayer HealthCare, West Haven, CT)

we describe provides a test bed for implementing potential solutions.

Limitations of the opt-out approach. Because of date shifting we adopted while generating the SD, events tied to specific dates (e.g., studies of seasonal allergies) cannot be evaluated.

A concern raised during the design was that if specific alleles linked to a disease are discovered in a de-identified sample, there is no way to inform the patients who have contributed those samples. On the other hand, there are only a few genotype–phenotype relationships that are sufficiently well understood to warrant clinical intervention on the basis of information from a genotyping assay; when this occurs, such testing should become part of routine clinical care. In fact, progress in identifying and validating genotype–phenotype relationships worthy of clinical intervention has suffered,^{11,12,35} in part because of the lack of DNA-clinical resources large enough to provide sufficient statistical power to allow the study of complex diseases and pharmacogenomic trends involving matrices of genetic variants.

Because there is no recontact with the patient for acquiring additional samples, the collection of serum samples for proteomic analysis is not possible. Although this may become feasible in the future, issues such as sample handling and stability will need to be addressed. Similarly, very rare phenotypes, such as those associated with certain adverse drug reactions, are unlikely to be well represented in this resource. Targeted databases, often requiring participation by multiple centers with specific expertise or patient populations, may be required in order to address some of these issues;³⁵ any large resource like this cannot address all questions for all investigators.

The portion of the SD that is simplest to use in research is structured data such as laboratory values and diagnostic codes. The SD also includes extensive unstructured data such as patients' histories. Natural language processing and other tools will be required for exploiting these data fully. Our early results with gender typing and allele identification in sickle cell disease and age-related macular degeneration suggest that the resource, when populated with larger numbers of samples, will become useful for identifying complex genotype–phenotype relationships. Indeed, the extent to which the mining of EMR information can be carried out to ascertain “real-world” phenotypes represents a major challenge as well as an opportunity for those working in the field of clinical informatics.^{36–40} Further, plans are under way for future genome-wide association studies that will use this resource and characterize its overall reliability and validity in the context of EMR-derived phenotypes (<http://www.gwas.net>), but these are beyond the scope of this paper.

The development of this DNA databank project has required extensive institutional resources, including long-term investment in information technology and specific costs such as an expanded DNA extraction and storage capability, review and implementation of new policies around the “consent to treatment” document, and support for community engagement and extensive internal and external review.

Although notification methods continue to be enhanced, we expect that a small minority of patients who would have chosen to opt out might not have done so because they overlooked the option box provided to them.

We describe here a complex plan for the implementation of a large de-identified DNA databank linked to extensive de-identified medical records. We believe this model provides advantages over an opt-in (consent) model, in not only scalability and richness of content, but also in terms of protection of genetic information. The development of the resource has required an institutional commitment to continuous investment in clinical genomics and biomedical informatics, both as clinical resources and investigational priorities. Further, the SD and DNA biobank resources—derived from actual EMRs—provide a test bed for the development and evaluation of new informatics and de-identification tools that will ultimately be required for implementing the vision of genome-based personalized medicine as part of the health-care system.

METHODS

Resource design. Over the past decade Vanderbilt University Medical Center, through its Department of Biomedical Informatics, has developed a comprehensive EMR system that covers all inpatient and outpatient data entry in the health system, including labs, drug ordering, and diagnostic imaging.^{22–27}

The Vanderbilt EMR is a state-of-the-art clinical and research resource that includes >1.4 million records and provides a platform for the development of tools, such as natural language processing approaches, to optimally mine structured data and unstructured (free-form) text in the medical records.

The major components of the Vanderbilt DNA databank initiative were developed on the basis of this resource, and these are presented in **Figure 4**. These include acquisition and de-identification of discarded blood samples and the creation of a “mirror image” of the entire EMR, termed the SD, which is similarly de-identified but retains a link to the DNA samples. In order to accomplish the goal of linking the clinical and DNA information in a de-identified fashion, the medical record number that labels each sample and each entry in the EMR is replaced with a research unique identifier (RUI) generated by the secure hash algorithm (SHA-512) developed by the National Security Agency of the US Federal Government (<http://www.nsa.gov/about/index.cfm>). SHA-512 is a publicly available hash function, an algorithm that produces a string of 128 characters that is unique to a particular input; in other words, it will always generate the same output (RUI, in this case) given the same input (medical record number, in this case). It is important to note that the original input (the medical record number) cannot feasibly be regenerated from the hash output (the RUI).⁴¹ No link is maintained between the RUI and the medical record number.

Regulatory and ethics review. The approach adopted is based on the guideline issued on 10 August 2004 by the Office for Human Research Protections (the branch of the US Department of Health and Human Services that is charged with protecting volunteers in research that is conducted or supported by the department). The document (<http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf>) concludes that use in research of discarded samples that are de-identified and not readily re-identified does not involve human subjects and so is not subject to the federal regulations for human subjects research (45 CFR 46, also called the Common Rule).

The design process is illustrated in **Figure 5**. The project was reviewed at multiple levels, including the IRB, ethics committee, community board, legal department, and Office for Human Research Protections. The Vanderbilt DNA databank is compliant with the regulations relating

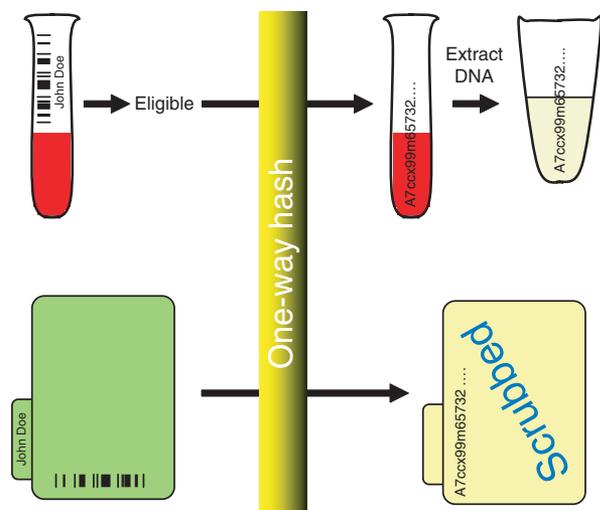


Figure 4 Mechanism for linking DNA samples and patient-related information in a de-identified fashion. The approach depends on the use of a one-way hash, an algorithm that always generates the same 128-character code (the research unique identifier, RUI) when the same medical record number is used as input. The medical record number on barcoded blood samples that are about to be discarded is scanned, eligible samples are relabeled with the RUI, and DNA is extracted and stored. The medical record number in each patient's record is replaced by the RUI, and the record is de-identified to create the synthetic derivative described in the text.

to the protection of human subjects as set forth by the Office for Human Research Protections and the Vanderbilt IRB. In addition, given the seminal nature of the program and the multitude of ethical domains affected, the protection model includes ongoing institutional and IRB oversight; evaluation by the Medical Center's Ethics committee; and the establishment of Ethics, Scientific, and Community Advisory Boards.

The review of the design among the various boards was iterative: after initial review and recommendations, the overall project, its components, and the results of preliminary validation studies were re-reviewed, leading to further recommendations for program adjustment. The studies described here, initiated on the basis of these reviews, evaluated the efficacy of sample handling and de-identification procedures and explored community response. Each was separately reviewed and approved by the IRB.

Interactions with the community. Three projects were conducted to gauge the community's reaction to a planned biobank: (i) a series of focus groups with lay members of the community (patients and nonpatients of varying socioeconomic, racial, and ethnic backgrounds); (ii) a mail survey consisting of a pretested, 38-item questionnaire sent to a random sample of 5,000 inpatients, outpatients, and emergency department patients; and (iii) an evaluation of the impact of placing posters describing the databank at blood-drawing stations.

Sample handling and interface with the SD. Clinical laboratories at Vanderbilt University Medical Center routinely retain blood samples (generally for ~3 days) before discarding them. Programs were developed to screen blood samples, obtained for routine clinical care and about to be discarded, for eligibility. The samples that are included are immediately transferred to tubes and relabeled with the 128-character RUI. Exclusion criteria are listed in [Table 2](#). Duplicate blood samples are excluded by executing the hash function in real time on an individual sample that meets all other inclusion criteria. At the time the hash value is generated, it is checked against the table of existing values. If the identical string of characters exists, the sample is deemed a duplicate and blocked from acceptance. A proportion of the eligible records (currently, ~50%) is randomly excluded so as to further ensure that it is

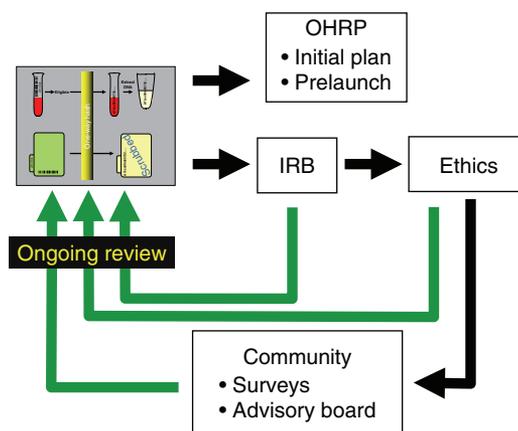


Figure 5 Program review. The program plan was reviewed by the Office for Human Research Protections (OHRP) and the Institutional Review Board (IRB). The IRB recommended further review from the standpoint of ethics, and the Ethics Review recommendations included the formation of a Community Advisory Board. The IRB, Ethics, and Community reviews resulted in program revisions, and these are ongoing.

Table 2 Inclusion and exclusion criteria

Manual exclusions
Poor quality
Insufficient blood volume
Automated exclusions
Minors
No signed consent to treatment form (includes the emergency department)
Samples from individuals who opted out
Non-Vanderbilt samples
Duplicate samples
A percentage randomly selected

never known which patients' samples (of those that could be included) are actually in the databank.

Two tests of the sample handling procedures and their interface with the SD were conducted. In the first, a set of 384 randomly selected samples that met the acceptance criteria were processed and genotyped for a gender-specific marker, using the Taqman amelogenin insertion/deletion polymorphism assay. The second test of sample handling involved genotyping a set of 24 samples with a diagnosis of sickle cell disease based on the medical record and 33 samples with a diagnosis of age-related macular degeneration and an associated common complement H susceptibility⁴² nested in a test collection of 243 control samples from patients who did not have either condition. The hemoglobin gene (*HBB*) was resequenced from 200 base pairs upstream from the start codon to 300 base pairs into the coding sequence, a segment that includes the majority of sickle cell mutations. The age-related macular degeneration complement H susceptibility allele (rs1061170) was assayed using a Taqman assay on an ABI 7900HT instrument (Applied Biosystems).

The SD. The SD is a database containing all clinical information in the EMR and its associated entry-order relational database but stripped of personal identifiers. The clinical systems have been in use since the mid-1990s and have been undergoing further development. They contain ~120 gigabytes (not including images) of information and >300,000,000 observations relating to 1.4 million subjects. On any given workday, there are between 4,200 and 5,200 concurrent users of the EMR, who

are health-care providers such as doctors and nurses. Both structured (e.g., name-value pairs of laboratory tests, ICD9/10 diagnosis codes) and unstructured but searchable data (e.g., narrative progress and procedure notes) are included.

The SD can be used for searching and aggregating sets of patients for genomic analysis or as a stand-alone clinical research resource. The SD was generated by applying multiple and iterative processing steps to the data in the EMR system. The first step was to replace the medical record number with the 128-character RUI, using the one-way hash function described earlier. The SD is updated regularly, relying on the constancy of the RUI (i.e., that the same input will always be mapped to the same hash output). In this way, as new entries are accumulated daily in the clinical information system, the scrubbing process is applied to the new additions and they are linked to the existing information within the EMR-derived SD record, thereby updating the SD.

Additionally, commercially available software (DE-ID, DE-ID Data Corp) was supplemented with preprocessing and postprocessing to “scrub” medical records of personal identifiers. The de-identification methodology is based primarily on the removal of the fields that are specified in Section 164.514 of the HIPAA privacy rule. These include:

1. Names;
2. All geographic subdivisions smaller than a State;
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages >89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of ≥ 90 years of age;
4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web universal resource locators;
15. Internet protocol address numbers;
16. Biometric identifiers, including finger and voice prints; and
17. Full face photographic images and any comparable images.

With the removal of those items the data are said to be de-identified; however, de-identification is not synonymous with anonymization. Anonymization is generally considered a binary process: samples are anonymized or are not. For the DNA databank, the outcome of de-identification is not binary; rather, it is continuous and recognizes that biological data are so inherently rich in attributes that the re-identification potential is never zero. Accordingly, a data use agreement has been put in place that permits only approved queries and data analyses, and specifically prohibits attempts at re-identification at the risk of institutional sanctions.

The system also shifts all dates and replaces each of the remaining HIPAA safe harbor provision identifiers with corresponding tags. All dates in the EMR are shifted 1–364 days into the past; the shift is different across records but constant within the records of each patient, thereby allowing temporal analyses such as the development of adverse effects after a drug. The initial SD was generated in the fall of 2006 and is currently updated regularly as described earlier.

The efficacy of this “scrubbing” process was initially evaluated by examining prescrubbed and postscrubbed records with an average size of 915 kilobytes. The instances of under- and overscrubbing were identified by reviewing manual charts, and the algorithms were modified accordingly. The process was iterative until an undermarking rate of <0.1% for complete HIPAA identifiers was achieved as assessed by a manual comparison between the original and the scrubbed records.

ACKNOWLEDGMENTS

The development of the program described here would not have been possible without inputs from the DNA Core Resource (Cara Sutcliffe and Marshall Summar), the Center for Human Genetics Research (Jonathan Haines, Marylyn Ritchie, Dana Crawford), members of the Department of Bioinformatics (Sunny Wang, Randy Carnevale, and Jacob Weiss), and members of the Medical Center’s Boards for Ethics, Community Advisory, and Operations Oversight. The design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, and approval of the manuscript were supported by institutional funding and by the Vanderbilt CTSA grant 1UL1 RR024975-01 from NCCRR/NIH. Dan Roden had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

CONFLICT OF INTEREST

The authors declared no conflict of interest.

© 2008 American Society for Clinical Pharmacology and Therapeutics

1. Garrod, A.E. *Inborn Errors of Metabolism* 2nd ed. (Henry Frowde and Hodder Stoughton, London, 1923).
2. Evans, D.A., Manley, K.A. & McKusick, V.A. Genetic control of isoniazid metabolism in man. *Br. Med. J.* **2**, 485–491 (1960).
3. Vesell, E.S. & Page, J.G. Genetic control of dicumarol levels in man. *J. Clin. Invest.* **47**, 2657–2663 (1968).
4. Forbat, A., Lehmann, H. & Silk, E. Prolonged apnea following injection of succinylcholine. *Lancet* **2**, 1067–1068 (1953).
5. Motulsky, A.G. Drug reactions enzymes and biochemical genetics. *J. Am. Med. Assoc.* **165**, 835–837 (1957).
6. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
7. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
8. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
9. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
10. Wellcome Trust Case Control Consortium. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39**, 1329–1337 (2007).
11. Giacomini, K.M. *et al.* The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clin. Pharmacol. Ther.* **81**, 328–345 (2007).
12. Roden, D.M. *et al.* Pharmacogenomics: challenges and opportunities. *Ann. Intern. Med.* **145**, 749–757 (2006).
13. Burke, W. & Psaty, B.M. Personalized medicine in the era of genomics. *JAMA* **298**, 1682–1684 (2007).
14. Compton, C. Getting to personalized cancer medicine: taking out the garbage. *Cancer* **110**, 1641–1643 (2007).
15. Gulcher, J. & Stefansson, K. Population genomics: laying the groundwork for genetic disease modeling and targeting. *Clin. Chem. Lab. Med.* **36**, 523–527 (1998).
16. Ollier, W., Sprosen, T. & Peakman, T. UK Biobank: from concept to reality. *Pharmacogenomics* **6**, 639–646 (2005).
17. Wilke, R.A. *et al.* Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clin. Med. Res.* **5**, 1–7 (2007).
18. Hinrichsen, V.L., Kruskal, B., O’Brien, M.A., Lieu, T.A. & Platt, R. Using electronic medical records to enhance detection and reporting of vaccine adverse events. *J. Am. Med. Inform. Assoc.* **14**, 731–735 (2007).
19. Lieu, T.A. *et al.* Real-time vaccine safety surveillance for the early detection of adverse events. *Med. Care* **45**, S89–S95 (2007).

20. Pulley, J.M., Brace, M.M., Bernard, G.R. & Masys, D.R. Attitudes and perceptions of patients towards methods of establishing a DNA biobank. *Cell Tissue Bank*. **9**, 55–65 (2008).
21. Pulley, J.M., Brace, M., Bernard, G.R. & Masys, D. Evaluation of the effectiveness of posters to provide information to patients about a DNA database and their opportunity to opt out. *Cell Tissue Bank*. **8**, 233–241 (2007).
22. Giuse, D.A., Giuse, N.B. & Miller, R.A. Evaluation of long-term maintenance of a large medical knowledge base. *J. Am. Med. Inform. Assoc.* **2**, 297–306 (1995).
23. Neilson, E.G. *et al.* The impact of peer management on test-ordering behavior. *Ann. Intern. Med.* **141**, 196–204 (2004).
24. Miller, R.A., Waitman, L.R., Chen, S. & Rosenbloom, S.T. The anatomy of decision support during inpatient care provider order entry (CPOE): empirical observations from a decade of CPOE experience at Vanderbilt. *J. Biomed. Inform.* **38**, 469–485 (2005).
25. Jirjis, J., Weiss, J.B., Giuse, D. & Rosenbloom, S.T. A framework for clinical communication supporting healthcare delivery. *AMIA Annu. Symp. Proc.* 375–379 (2005).
26. Butler, J. *et al.* Improved compliance with quality measures at hospital discharge with a computerized physician order entry system. *Am. Heart J.* **151**, 643–653 (2006).
27. Boord, J.B. *et al.* Computer-based insulin infusion protocol improves glycemia control over manual protocol. *J. Am. Med. Inform. Assoc.* **14**, 278–287 (2007).
28. Huang, N., Shih, S.F., Chang, H.Y. & Chou, Y.J. Record linkage research and informed consent: who consents? *BMC Health Serv. Res.* **7**, 18 (2007).
29. Buckley, B., Murphy, A.W., Byrne, M. & Glynn, L. Selection bias resulting from the requirement for prior consent in observational research: a community cohort of people with ischaemic heart disease. *Heart* **93**, 1116–1120 (2007).
30. Hewison, J. & Haines, A. Overcoming barriers to recruitment in health research. *BMJ* **333**, 300–302 (2006).
31. Junghans, C., Feder, G., Hemingway, H., Timmis, A. & Jones, M. Recruiting patients to medical research: double blind randomised trial of “opt-in” versus “opt-out” strategies. *BMJ* **331**, 940 (2005).
32. Malin, B.A. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *J. Am. Med. Inform. Assoc.* **12**, 28–34 (2005).
33. Lowrance, W.W. & Collins, F.S. Ethics. Identifiability in genomic research. *Science* **317**, 600–602 (2007).
34. Clayton, E.W. Ethical, legal, and social implications of genomic medicine. *N. Engl. J. Med.* **349**, 562–569 (2003).
35. Giacomini, K.M., Krauss, R.M., Roden, D.M., Eichelbaum, M., Hayden, M.R. & Nakamura, Y. When good drugs go bad. *Nature* **446**, 975–977 (2007).
36. Chute, C.G. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. *Proc. AMIA Symp.* 165–169 (2002).
37. Sax, U. & Schmidt, S. Integration of genomic data in electronic health records—opportunities and dilemmas. *Methods Inf. Med.* **44**, 546–550 (2005).
38. Meystre, S. & Haug, P.J. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J. Biomed. Inform.* **39**, 589–599 (2006).
39. Nuzzo, A., Segagni, D., Milani, G., Rognoni, C. & Bellazzi, R. A dynamic query system for supporting phenotype mining in genetic studies. *Medinfo*. **12**, 1275–1279 (2007).
40. Hoffman, M.A. The genome-enabled electronic medical record. *J. Biomed. Inform.* **40**, 44–46 (2007).
41. Computer Security Resource Center, National Institute of Standards and Technology. Federal Information Processing Standards Publication 180-2: Secure Hash Standard <<http://csrc.nist.gov/publications/fips/fips180-2/fips180-2.pdf>> Accessed 1 November 2007.
42. Haines, J.L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).