HARVARD UNIVERSITY

Formalized Procedure for HarvardX Data Requests

Dustin Tingley, Andrew Ho, Brooke Pulitzer, and Yigal Rosen
Vice Provost for Advances in Learning Research

February 2016

The increased interest in research on learning science, online learning, and evidence-based practice in higher education requires formalized processes surrounding Harvard data sources. As part of the Vice Provost for Advances in Learning (VPAL) mission, the VPAL-Research (VPAL-R) group has established a process whereby researchers – both internal and external to Harvard University – may conduct experiments and analyze data from HarvardX courses. This process builds on the foundation of previous policies (developed under the auspices of the then HarvardX Research Committee) and studies by developing generic and customized data products that can be applied toward and adapted for a variety of research endeavors toward the benefit of students, instructors, researchers, and society at large. With these goals in mind, VPAL-R – in conjunction with the Committee on the Use of Human Subjects (CUHS), Office of the General Counsel (OGC), Office of the Vice Provost for Research (OVPR), Office for Sponsored Programs (OSP), and a number of individuals across the University – have established a formal procedure and workflow for requesting, reviewing, and providing researchers with data from HarvardX courses (outlined herein). This protocol replaces any earlier iteration surrounding the request of HarvardX data.

Figure 1 summarizes the procedure for requesting, reviewing and providing researchers with HarvardX data. First, a data request form will be submitted to record all external/internal data requests via: https://harvard.az1.qualtrics.com/SE/?SID=SV_ag6DIW9bQTGHsZT
Yigal Rosen (VPAL-R) will actively monitor and review requests on a weekly basis and provide researchers with guidance on the IRB process, if needed. Upon approval of the data request, researchers will be guided to follow one of the three paths depending on the type of research project (e.g., internal/external to Harvard) and data requested:

a. FERPA de-identified data
b. Restricted identifier data
c. FERPA identified data

All research proposals involving FERPA identified and/or restricted identifier data will be vetted at several levels, including VPAL-R (which will check to ensure that ethics training and IRB approval requirements have been met in the intake form), the CUHS (where FERPA identified data is concerned), and the OVPR. Note that, at this time, any external (non-Harvard) researchers wishing to access restricted identifier data must partner with a Harvard faculty

collaborator, or "sponsor." Also note that a pared-down version of the requisite Data Usage Agreement (DUA) will be employed for internal faculty researchers requesting FERPA identified or restricted identifier data, who are not doing so in collaboration with external researchers.

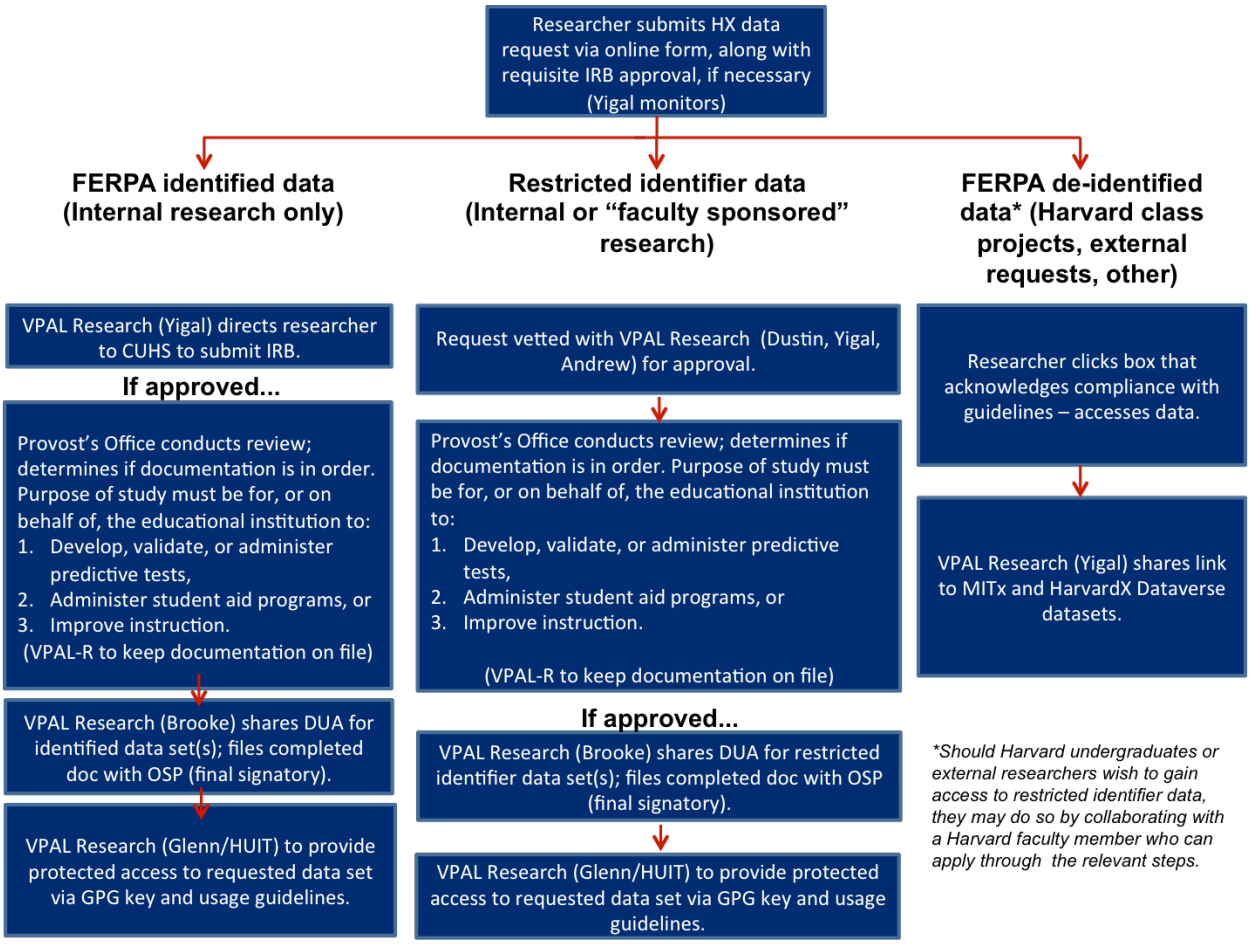


Figure 1. Procedure for requesting, reviewing, and providing researchers with HarvardX data

Class projects. Harvard faculty-sponsored class projects that involve HarvardX restricted identifier data (as opposed to FERPA de-identified data) will be required to apply for IRB approval (template/sample attached as Appendix). The faculty sponsor will be asked to provide course-project specific objectives and describe his/her experience with the proposed research procedures. Additionally, the faculty sponsor and all students involved in the project will be asked to complete CITI training and sign a DUA for restricted identifier data prior to receiving access.

Types of datasets. The VPAL-R team is able to provide three types of datasets, referred to as the "Person-Course-Survey," "Person-Course-Day," and "Log Data." A list of the fields contained in these datasets is documented in the Harvard VPAL Private Datasets. http://vpal.harvard.edu/files/vpl/files/vpal_private_datasets.pdf

Delivery. Prior to delivery, VPAL-R will request that the principal investigator follow the process of creating keys for encryption and decryption. This one-time process involves the requestor's creating a pair of keys using GNU Privacy Guard (GnuPG or GPG): 1) the public key, which is sent to the Harvard VPAL Research Team in order to encrypt data, and 2) the private key, which the requestor then uses to decrypt files that have been encrypted with that public key. This process involves the requestor installing a cryptographic application locally on their machine and supplying his/her official academic institution email address along with a secret passphrase.

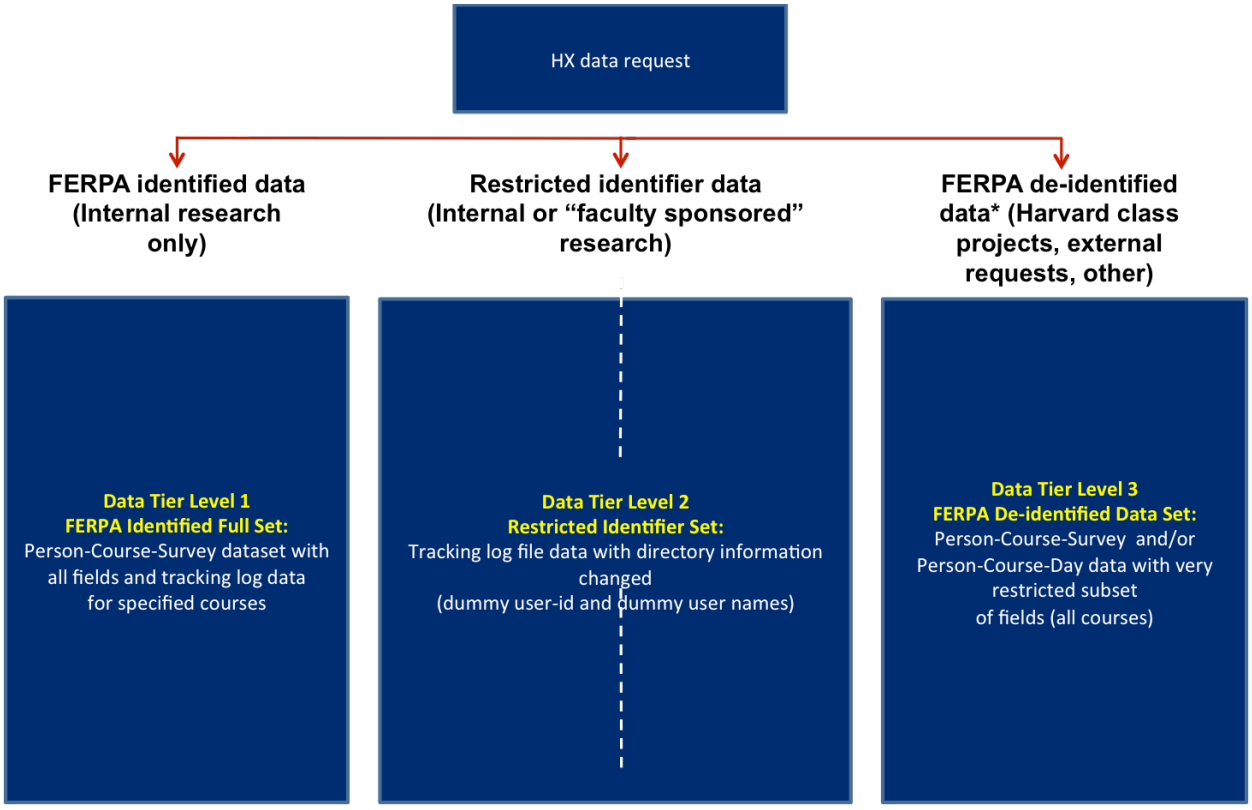


Figure 2. Types of HarvardX data requests

HARVARD
UNIVERSITY

Figure 2 illustrates data tiers and available data sets for each tier:

- Data Tier Level 1 (FERPA Identified Full Set) includes Person-Course-Survey dataset with all fields and tracking log data for specified courses. Once approved through VPAL-R, the data will be released to the principal investigator through a secure project space or delivery through the GPG public/private key pair download.
- Data Tier Level 2 (Restricted Identifier Set) includes log file data with directory information changed (dummy user-id and dummy user names). Once approved through VPAL-R, data will be released to the principal investigator through the GPG public/private key pair download.
- Data Tier Level 3 (FERPA De-identified Data Set) includes Person-Course-Survey and/or Person-Course-Day data with very restricted subset of fields (all courses). Once approved through VPAL-R, the data will be released publicly through Harvard VPAL Dataverse (https://dataverse.harvard.edu/dataverse/VPAL), along with configuration-controlled documentation synchronized with each dataset.

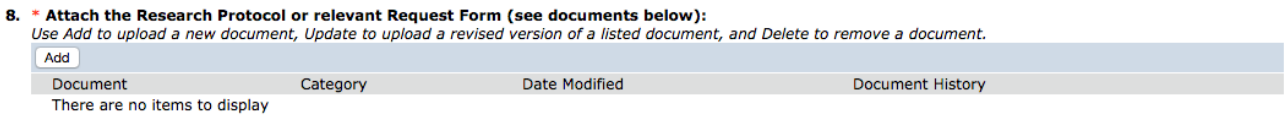
We invite researchers to contact Yigal Rosen for further information on IRB procedure and support. Formal requests for research collaboration and data requests should be submitted via: https://harvard.az1.qualtrics.com/SE/?SID=SV_ag6DIW9bQTGHsZT

HARVARD UNIVERSITY

Appendix

IRB Template for Harvard Faculty-Sponsored Class Project Request for HarvardX Restricted Identifier Data

Instructions: Complete all of the sections below (type an x in a Yes/No box or provide an answer). The wording of the questions includes some explanatory material designed for typical CUHS research; more detailed instructions may be found in the Template Guide. If your study is more than minimal risk, you **must** consult the Template Guide for definitions and additional required material.
Then go to our online submission system ESTR (https://irb.harvard.edu/), sign in with your HUID, and click on Create New Study. Fill out the online pages and upload this Research Protocol in the appropriate place:


8. * **Attach the Research Protocol or relevant Request Form (see documents below):**
Use Add to upload a new document, Update to upload a revised version of a listed document, and Delete to remove a document.
Add
Document | Category | Date Modified | Document History
There are no items to display

GENERAL INFORMATION	
Protocol # (if assigned):	**Version Number or Date:**
Principal Investigator Name: [insert faculty sponsor's name here]	
[X] **Faculty** [ ] **Graduate student** [ ] **Post-doc** [ ] **Undergraduate** [ ] **Extension school student** [ ] **Junior Fellow** [ ] **Staff** [ ] **Visiting Scholar** [ ] **Other** (specify)**:**	
Faculty Sponsor (if Principal Investigator is not faculty)**:**	
Other Advisor Name (if applicable)**:**	
[ ] **Supervising lecturer** [ ] **Instructor** [ ] **Graduate student** [ ] **Thesis advisor** [ ] **Other** (specify)**:**	
Protocol Title: Course project: [insert project title here]	

1. Background

1.1. Provide the scientific background, rationale for the study, and importance in adding to existing knowledge.

HarvardX is the Harvard University component of edX, a joint venture with the Massachusetts Institute of Technology (MIT) for open online course development and administration. This course project proposes to analyze user data produced from HarvardX courses SW12x – SW12.10x (ChinaX Parts 1 – 10 version 1) between the dates of 5/29/13-11/9/15.

Even with more universities offering massive open online courses (MOOCS), they continue to be a new venture at selective institutions of higher education. The characteristics of the students, their interactions online (both with the course content and each other), and factors that contribute to their academic success in MOOCs are still in the early stages of research. The data gathered from widely

accessible HarvardX courses represent a unique opportunity to answer major questions about open-access online learning from elite institutions.

In this application, we describe a highly feasible, high priority study that will follow-up on the important and successful work of the Year Two Course Reports.

The findings hold immense potential for informing the design of current and in-development HarvardX courses, the number of which is likely to increase dramatically over the next few years.

[insert course-project specific objectives]

2. Study Design

2.1. Provide a thorough description of all study procedures.

We will use descriptive statistics, exploratory data analysis and other standard statistical techniques to better describe and understand learning processes and outcomes through the courses.

When students register for edX, initial questions about their name, educational background, gender, birth year, mailing address, and motivation for registration are asked. Students subsequently view lectures, post on discussion forums, and complete course assessments and projects. This represents standard participation in HarvardX courses.

This range of activities falls under exemption category 1, research conducted in normal education settings. In high quality classrooms, teachers provide educational resources and use tools to learn about their students' interests, motivations, experience, preparedness, and background.

If any of the above activities do not fall in exemption category 1, we believe they fall under exemption category 2. There may be a small number of questions on embedded surveys that are less likely to be asked by teachers in typical educational settings, particularly demographic questions such as "level of parental education." We record identifying information separately from survey responses, and we do not anticipate that any harms could come to students should their survey responses be inadvertently released. (These surveys and associated documents are already part of IRB 1376-06, Surveys and Student Activity in HarvardX).

This project will use data Tier Level 2 - Restricted Identifier Set: Tracking log file data with directory information changed (dummy user-id and dummy user names). This data set will contain time series information along with dummy user-id/usernames, clickstream events, activity and the respective course id.

2.2. Indicate the duration of a participant's involvement.

Data collection will be ongoing with the pace of standard course interactions, including registration, completion of assignments, participation in online discussion groups, and completion of final exams. Brief surveys are incorporated regularly into problem sets. These surveys are again primarily designed for course improvement, for example, asking about the difficulty and clarity of the assignments, or what

students feel they have learned.

2.3. Indicate the estimated number of participants, by subgroup if applicable.

Overall, 103,427 participants are included in the dataset.

2.4. List inclusion and exclusion criteria and describe any screening process.

None

2.5. Does the study involve (a) deception (providing false information) or (b) incomplete disclosure (withholding information about some or all aspect of the research purpose or procedures in order to maintain the scientific integrity of the study)?
[X] No [ ] Yes: If yes, explain the rationale and plans for protecting participants (e.g., debriefing).
Be sure to attach any debriefing materials to the "Supporting Documents" webpage.

3. Recruitment Methods

3.1. Will potential participants be provided with information about the study?
[X] No: If no, skip to 4.1.
[ ] Yes: If yes, indicate how, when, where, and by whom participants will be recruited.
If you are recruiting from a study pool, describe how you meet their requirements (see Template Guide).

3.2. Are there any materials that will be used to recruit participants (e.g., emails, posters, oral scripts)?
[X] No [ ] Yes: If yes, list the materials by document name here, and be sure to attach final copies to the "Consent and Recruitment Materials" webpage.

3.3. Will participants receive reimbursement or compensation in the form of money, gifts, incentives, or raffles?
[X] No [ ] Yes: If yes, specify the amount, method and timing of disbursement.
See Template Guide for specific information on payments and a link to the Harvard University Financial Policy on Human Subject Payments.

There is no payment and no inducement associated with this research above and beyond the inducement that the course itself represents.

4. Study Setting

4.1. Is any of the research conducted outside the United States?
[X] No [ ] Yes: If yes, describe how you are ensuring that the research is appropriate considering local laws, regulations, and customs.
This should be either a formal review by a local ethics board, Ministry of Health, etc., or a

statement that a formal review is not required along with your source of information that the proposed research is in accordance with local laws, regulations, and customs.

4.2. Are there any permissions that must be obtained from cooperating institutions, community leaders, government officials, or other individuals, including approval from an IRB or research ethics committee?
[X] No [ ] Yes: If yes, list the permission(s) by document name and be sure to attach copies to the "Supporting Documents" webpage.

5. Available Resources

5.1. Describe the experience of the investigator with the proposed research procedures and population.

[insert faculty sponsor's research experience here]

5.2. Are there any additional study team members whose role in the research require special qualifications in addition to ethics training (e.g., licensed clinical psychologist)?
[ ] No [X] Yes: If yes, describe.

Students that will be involved in this course project will be requested to sign on the DUA (see Appendix A)

5.3. Are provisions needed for medical and/or psychosocial support resources (e.g., in the event of research-related distress or incidental findings)?
[X] No [ ] Yes: If yes, describe the provisions and their availability.

6. Vulnerable Populations

6.1. Are there any potentially vulnerable populations or individuals (minors, pregnant women, human fetuses, neonates, prisoners, economically disadvantaged, employees or students of the investigator, cognitively impaired, etc.) proposed for involvement in the research?
[ ] No [X] Yes: If yes, identify all vulnerable populations and describe proposed safeguards to protect their rights and welfare.

Although veriety of learners may be expected to enroll in HarvardX courses, we anticipate that there will be no more than minimal risk associated with their participation given the security measures in place to prevent the release of identifying information, and that data are collected for the primary purpose of improving and informing instructional practices.

7. Consent Process

7.1. Will participants be asked to agree to be in the study?
[ ] No: If no, explain why not, then skip to 8.1.
[X] Yes: If yes, describe the consenting process.
If the study includes minors or others who cannot consent for themselves, describe how you

will obtain their assent and the permission of their parent or guardian. Be sure to attach copies of appropriate documents to the "Consent and Recruitment Materials" webpage.

Participating students consent to a Terms of Service agreement and attendant privacy policy available at https://www.edx.org/edx-terms-service.

7.2. Will the consenting process involve obtaining a signature?
☐ **Yes** ☒ **No: If no, explain why not.**
The requirement to obtain a participant's signature can usually be waived by CUHS for minimal risk research, see Template Guide.

7.3. Will participants be offered a copy of the consenting information?
☒ **Yes** ☐ **No: If no, explain why not.**

Terms of Use are always accessible from the edX.org website, and links will be provided in solitications.

7.4. Are you recruiting any participants who are not fluent in English?
☐ **No** ☐ **Yes: If yes, describe provisions for communicating information needed for consent.**

7.5. Are you audio or video recording any participants?
☒ **No** ☐ **Yes: If yes, describe provisions for notifying participants of this recording.**

8. Risks

8.1. Are there any reasonably foreseeable risks or discomforts to participants and/or groups/communities?
☒ **No** ☐ **Yes: If yes, describe the risks and outline proposed provisions to minimize risk.**
Risks may be physical, psychological, social, legal, and/or economic. If risks are more than minimal, there are additional questions you must answer, see Template Guide.

The data that will analyze have been contributed by participants voluntarily, for the purpose of informing and improving open online instruction. All of the data collected is standard to participation in the HarvardX classes and part of the standard feedback and evaluation mechanism for course development and improvement. We anticipate no risks to the collection and analysis of course interaction data collected in this manner that requires no intervention in courses. For survey data, there is only the minor inconvenience that accompanies online surveys, which participants can end at any time by simply closing their browser.

9. Data Confidentiality

9.1. Which category of information best describes the data you will be recording?
Refer to Template Guide for additional information.
☐ **You will not collect any direct or indirect individual identifiers (Level 1). Explain.**

☐ **Participants will be told that their data will be made public (Level 1). Explain.**

☒ **The data will be identifiable but not sensitive (sensitive information could be damaging to the participants if revealed), and participants will be told that their data will not be shared outside the research team (Level 2). Explain.**

This project will use data Tier Level 2 - Restricted Identifier Set: Tracking log file data with directory information changed (dummy user-id and dummy user names). This data set will contain time series information along with dummy user-id/usernames, clickstream events, activity and the respective course id.

☐ **The data will be identifiable and sensitive (Level 3, 4, or 5, depending on the degree of sensitivity). Describe how sensitive the information is and the protections you have developed in consultation with the appropriate IT resource.**

9.2. Where will the data be stored?

The event log data are initially stored primarily by edX. These data are then encrypted and securely transmitted to the VPAL Research Team, under the supervision of Jim Waldo (CTO for Harvard). These data are then maintained on secure servers with restricted access at the Institute for Quantitative Social Science.

9.3. Describe i) plans for any transmission of identifiable data; ii) how long and with what protections identifiable data will be stored; and iii) plans for the data at the end of the storage period (how it will be destroyed, or if it will be returned to the data provider).

Data will be accessible to only the principal investigator and his students. In addition, no information will be reported that would allow for the identification of individual students or professors.

Jim Waldo, Harvard's Chief Technology Officer, manages access to an IQSS server with course log data and demographic data from the edX site registration survey.

9.4. Indicate how research team members, other collaborators, or other researchers are permitted access to information about study participants.

VPAL Research Team implements the following secure data sharing process in course-projects:
1. EdX / Qualtrics Data
Download data from edX Amazon Server [to local machine] and Qualtrics [to IQSS]
Decrypt Hx course database extracts and event log data [on local machine]
2. Extract / Transform / Load
Connect to IQSS project space "ci3_jwaldo" from local machine, upload and archive raw database

extracts to IQSS project space “ci3_jwaldo” in weekly folder for each respective course
Execute python extraction/transform/load scripts on IQSS, and load data to Cloud data warehouse
3. Research Data Sets
Execute python scripts to generate Person-Course-Survey research data set on Cloud data warehouse
Extract research data from Cloud data warehouse to IQSS project space “ci3_jwaldo”
4. Prepare and deliver data to researchers
Encrypt data with public key/private key pair, or
Provide access to “ci3_jwaldo”

10. Benefits

10.1. Describe any potential benefits to study participants and to society.

Participation in this course project will benefit future learners through improved targeting of instructional resources to instructional needs. Students will gain access to a Harvard education for free.

11. Participant Privacy

11.1. Describe provisions to protect participants’ privacy (their ability to control access to information about themselves or their person, e.g., the use of a private interview room) and to minimize the intrusiveness of study questions or procedures.

Participants complete all course materials and surveys from their own computing resources, and they have the power to stop participation at any time by simply closing their browser.

12. Sharing Study Results

12.1. Is there a plan to share study results with individual participants and/or the participant community?

[X] No [ ] Yes: If yes, describe the plan.

13. Multi-site Study Management

13.1. Are one or more sites conducting this study in addition to sites overseen by the Harvard PI?

[X] No [ ] Yes: If yes, indicate whether there is a coordinating research site and describe plans for communication among sites regarding unanticipated problems involving risks to subjects or other individuals, interim results, protocol modifications, monitoring of data, etc.

14. Devices

14.1. Does this study involve the use of a device subject to FDA regulations?

[X] No: If no, skip to 15.1

[ ] Yes, and the device is being used according to its labeled indication: Skip to 15.1

[ ] Yes, and the device is an Investigational Device: Describe why this is a non-significant risk device study and why it qualifies either for an abbreviated IDE determination or for exemption from the IDE requirements.

HARVARD UNIVERSITY

15. HIPAA Privacy Protections

15.1. Are HIPAA privacy protections required? Mark Yes only if the investigator is at Harvard University Health Services or data will be obtained from a hospital, health center, or health insurance plan (see Template Guide).

[X] No [ ] Yes: If yes, either describe plans for obtaining authorization to access protected health information or provide the scientific or logistical rationale for a waiver of authorization or limited waiver of authorization request.

16. Establishing a Data or Specimen Bank

16.1. Does the study include establishing a repository for sharing data or specimens with other researchers? *This does not include contributing de-identified data to an existing repository.*

[X] No: If no, then there are no more questions.

[ ] Yes: If yes, identify what data or specimens will be collected and stored, and what information will be associated with the specimens.

16.2. Describe where and how long the data/specimens will be stored and whether participants' permission will be obtained to use the data/specimens in other future research projects.

16.3. Identify who may access and use data/specimens and how.

16.4. Will specimens and/or data be sent to research collaborators outside of Harvard?

[ ] No [ ] Yes: If yes, describe the plan, and be sure to attach copies of any agreements to the "Supporting Documents" webpage.

16.5. Will specimens and/or data be received from collaborators outside of Harvard?

[ ] No [ ] Yes: If yes, describe the plan, and be sure to attach copies of any agreements to the "Supporting Documents" webpage.