Appendix A. The NSRR technical and operational details

Metadata extraction pipeline. This pipeline involves the following five specific steps.

1. Capture metadata for the cohort on such aspects as What, Who, When, and Funding Source. Standard provenance information captured for each study includes "About," overview of "Core clinical data" and "Overnight PSG data," "Protocols and Manual of Procedures," "Montage and Sampling Rate Information Analysis," "PSG Sleep Variable Data Guide," "EEG Spectral Analysis," "Heart Rate Variability (HRV) Analysis," "Thoraco-Abdominal Asynchrony Analysis (TAA)," and links to related "Publications."

2. Integrate sleep and non-sleep data. Information about each study-subject's sleep, demographic and health characteristics/outcomes are exported from a source SAS dataset and imported to the Spout JSON data dictionary format. The Spout tool provides tests to assure that the data are fully described and that the data dictionary covers the full dataset. Spout also handles deployment of versioned datasets to the NSRR, and generates statistics, graphs, and metadata for each element of the dataset.

3. Convert PSGs to EDF files. PSG files are converted to EDF files using the set of EDF tools adapted from Physio-MIMI. Normalization of signal header and scrubbing of any embedded sensitive information (dates) in header fields also occurs. The integrity of the EDF files is checked by the Edfize tool before and after the EDF files are ported to the NSRR.

4. Apply signal analyses in batch. After signal attributes are normalized and artifact edited, a suite of signal analysis tools is applied to the signal files. This derived set of signal features are stored in a separate relational table linked to the study, with a unique identifier associated with each PSG/EDF file. A copy of an artifact-edited version of the PSG signal files is saved and

linked to the individual participant's study identifier. All PSGs and related clinical data are given a new unique, random identifier.

5. Map, curate, and index all study terms to allow quick web-based overview of the basic summary statistics associated with each term. The Spout tool is used to deploy updates to datasets and their corresponding data dictionaries to the NSRR. The NSRR also seamlessly archives older versions of data definitions and data for reference, reproducibility, and dataset history.

Available tools for data visualization and analysis. To promote data reuse, the NSRR web portal created a dedicated area (https://sleepdata.org/tools) for sharing tools developed by the NSRR team as well as contributed by the community. A select subset of tools is highlighted below.

- EDF Viewer (in Matlab) allows the user to open an EDF file and the corresponding sleep annotation file;
- EDF Browser is a multi-platform, universal viewer and toolbox that can be used to view continuous biomedical signals;
- EDF Editor and Translator facilitates the de-identification and standardization of signals by editing EDF files and translating sleep annotations generated in different vendor formats to sleep domain standardized terminology;
- SpectralTrainFig (in Matlab) is a pipeline for EEG spectral analysis. A graphical interface allows the user to input EDF files and annotations in XML format and returns the power spectral density of each EEG signal both on an epoch-by-epoch basis and summarized by sleep state. Coherence analysis between EEG leads can also be performed. The program includes an artifact detection step, based on delta and beta band

descriptors, and a template-subtraction algorithm for decontaminating the EEG signals from ECG artifact. The interface also allows the identification of sleep cycles, based on adapted Feinberg and Floyd criteria, and cycle-specific spectral analysis;

- XML Annotation Extractor (in R) allows conversion of data in XML files into a single, combined CSV file that lists sleep staging and scoring events;
- Data Chromatix (in Matlab) is a visualization method for continuous physiological signals which colorizes data in the current display window based on information about the system's past behavior, allowing to quickly recognize trends and variations;
- Multiscale Poincaré Plots (in Matlab) is a graphical tool to facilitate visualization and discovery of patterns in complex signals, such as interbeat interval time series;
- MSE-based EEG artifact detector (in Matlab) employs multiscale entropy analysis to detect noisy epochs on an EEG signal;
- ActiCircadian (in Matlab) performs parametric (cosinor) and non-parametric analysis for the extraction of rest-activity features from actigraphy data.

Data access and data protection. The NSRR offers two main levels of data access: online and offline. The x-search tool supports online data exploration, and the Altamira tool supports online data visualization. Such online tools are implemented as web applications in the underlying architecture. Offline tools include EDF Viewer, EDF Browser, EDF Editor and Translator, SpectralTrainFig, XML Annotation Extractor, Data Chromatix, Multiscale Poincaré Plots, MSE-based EEG artifact detector, and ActiCircadian. These offline tools are provided as executable programs and can be downloaded by users to perform data visualization and analysis activities using their local computational resources.

Data access at subject level is based on a user’s active Data Access and Use Agreement (DAUA). Each DAUA provides access to one or more datasets for a fixed amount of time (1 to 3 years). The DAUA provides a user individual access to all the de-identified data in a particular dataset. Users without a DAUA are however still able to see aggregate data collected on a dataset. The x-search interface leverages this by providing users who don’t have DAUAs the ability to query and view aggregate results of the underlying data.

Access to the NSRR web-interfaces is secured using HTTPS encryption from the users’ web browser to the web server. Communication between servers is managed through a scalable encryption algorithm layered on top of public/private key encryption. The algorithm interleaves symmetric encryption (for scalability) with asymmetric encryption (for security) to achieve a highly secure and rapid data-messaging framework. The de-identification process is automated so that it can be applied uniformly and consistently. All dates are scrubbed, with dates of events recorded in reference to study entry times. All other potentially identifiable data (e.g. extreme ages) are also removed.

Strategy for obtaining agreements for sharing data. One of the non-technical challenges is obtaining agreements to share data through the NSRR from the “owners” of individual study cohorts. Our first approach for populating data to the NSRR was to include data from studies that members of our team were directly involved with as investigators, and in fact, studies which were supported by our national Sleep Reading Center. Even with established relationships with each of the studies, however, there were several issues that needed to be addressed. In particular, many of the data sources are from studies that are governed by individual Steering Committees, each with varying perspective on data sharing. Potential concerns expressed while establishing the NSRR by potential partners included: ownership of data and authorship; input into decision making and governance regarding data use; privacy and compliance with regulatory agreements, including participant consent; and potential duplication of data across other repositories (such as BioLINCC or the study's own repository, which could result in several sources of data that could drift over time if each repository conducts unique edits). In addition, institutional concerns include sharing of original sleep records that contain limited Protected Health Information (dates).

The NSRR addressed such issues by: (1) Adopting flexibility in approaching data requests from each entity, particularly respecting requests to withhold sets of data (e.g., until sufficient time had passed when primary investigators had opportunity to publish data). (2) Collaborating with our institutions' Institutional Review Boards to ensure regulatory compliance and adherence with any restrictions in the original consent forms signed by study participants. (3) Establishing appropriate Data Use Agreements between institutions and establishing a Data Use Agreement between the NSRR and each individual user. (4) Inviting a representative from each study to participate in either the NSRR Steering Committee of a User Group as an opportunity to provide input for data governance. (5) Strictly removing PHI from shared data. (6) Coordinating with the originating site and other resources (e.g. BioLINCC) so that data elements are mapped or represented using common terms. Care was taken to avoid making changes in the data field labels (but annotating records that may be outliers or require edits before use) as well as to maintain the same participant identifiers to facilitate cross-linking of data across resources. Whenever possible, our data use agreements reflected those of BioLINCC, to ensure consistency. To support those needs without encumbering access, we developed well-controlled and monitored on-line procedures for reviewing and approving each application with on-line tools.

We created an online process for obtaining institutional approvals for individual data use agreements as well as tracking IRB status.

To encourage deposition of additional data sets, we created a document, available online at sleepdata.org, called “Sharing your own data on the NSRR.” This document outlined the process, expectations, and responsibilities of the contributor and the NSRR.

Hosting environment. The NSRR is deployed on Partners Research Information Science & Computing’s (RISC) Discovery Informatics Platform for Research (DIPR), a flexible and secure cloud computing infrastructure. The core of DIPR is two VMWare ESXi clusters that provide Vsphere management. The NSRR runs on four CentOS 7 Virtual Machines in the DIPR environment. Three servers run as dedicated application servers, with the fourth acting as a load balancer and proxy server. This architecture is designed specifically for zero downtime deployment. The servers also connect to the Research File Area which grants the 10 TB of storage that can be scaled as needed as data storage requirements increase. The NSRR web application is coded in Ruby on Rails 5.1 and runs on Phusion Passenger for NGINX web servers.

Because of the funded project scope, the NSRR did not pursue computation in the cloud for researchers. The architecture is designed to provide the data to the users, and to be able to maintain data access integrity between the user and the resource. The NSRR does facilitate and encourage the sharing of the analytic tools (through GitHub), but does not actually provide computational power to host the computation (using the “move computation to data” paradigm). Moving to a public cloud, or full integration with the NIH Data Commons, might be future directions to ensure sustainability. These are possible strategies that the NSRR team will be watching closely.

Appendix B. User feedback and distribution of downloaded data

The NSRR established an Academic User Group (AUG) in late 2013 and held an inaugural meeting in January 2014 prior to the official launch of the NSRR website. The AUG consists of key representatives from study groups with data on the NSRR, community users, and technical experts and was convened to provide input into the design and functionality of the resource. Prior to the official unveiling of the NSRR, feedback related to these aspects of the site was received through web demonstrations and surveys sent to the AUG. To further enhance development of the resource and promote interaction and collaboration with the user community, an Early Adopter Group (EAG) was formed in 2015. The primary goals of the EAG are to promote interaction and collaboration between the end users of the resource and the core development team, to help guide functionality and give input on posted data and tools, and to promote the sharing of ideas for future planning and resource development. An EAG meeting held in Boston, Massachusetts in October 2015 consisting of the NSRR core team and 12 invited guests resulted in 22 action items from the group that continue to guide resource development and community collaboration. The NSRR team seeks feedback from the user community through the discussion forum on the resource and the support@sleepdata.org email address, and via surveys sent to end users inquiring about usability and functionality of the site and available tools. A local NSRR Community Summit occurred on June 19, 2017 in Boston for the purposes of presenting the latest updates regarding data and tools in the NSRR, discussing future enhancements, and identifying opportunities for collaboration. An online User Forum (https://sleepdata.org/forum) provides an interactive, ongoing virtual space for the end users to interact with the NSRR team and each other on common questions, answers and topics of further interest.

Table A. The NSRR data download statistics

Dataset	**Files Downloaded**	**Data Downloaded**
Sleep Heart Health Study	4,475,072 files	50.0 TB
Childhood Adenotonsillectomy Trial	116,588 files	21.8 TB
Heart Biomarker Evaluation in Apnea Treatment	24,315 files	1.1 TB
Cleveland Family Study	99,557 files	8.2 TB
Study of Osteoporotic Fractures	18,268 files	709 GB
MrOS Sleep Study	368,467 files	29.8 TB
Cleveland Children's Sleep and Health Study	22,338 files	3.3 TB
Hispanic Community Health Study / Study of Latinos	309,737 files	388 GB
Honolulu-Asia Aging Study of Sleep Apnea	50 files	760 MB
Multi-Ethnic Study of Atherosclerosis	368,401 files	17.4 TB
TOTAL	5,802,793 files	133 TB