fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning

Florian Knoll, PhD* • Jure Zbontar, PhD* • Anuroop Sriram, MS • Matthew J. Muckley, PhD • Mary Bruno, BS • Aaron Defazio, PhD • Marc Parente, BS • Krzysztof J. Geras, PhD • Joe Katsnelson, BS • Hersh Chandarana, MD • Zizhao Zhang, PhD • Michal Drozdzalv, PhD • Adriana Romero, PhD • Michael Rabbat, PhD • Pascal Vincent, PhD • James Pinkerton, BS • Duo Wang, MS • Nafissa Yakubova, BS • Erich Owens, MS • C. Lawrence Zitnick, PhD • Michael P. Recht, MD • Daniel K. Sodickson, MD, PhD • Yvonne W. Lui, MD

From the Department of Radiology, NYU School of Medicine, 650 First Ave, New York, NY 10016 (F.K., M.J.M., M.B., M.P, K.J.G., J.K., H.C., D.W., N.Y., M.P.R., D.K.S., Y.W.L.); Department of Artificial Intelligence Research, Facebook, Menlo Park, Calif (J.Z., A.S., J.P, E.O., C.L.Z.); Department of Artificial Intelligence Research, Facebook, New York, NY (A.D.); Center for Data Science, New York University, New York, NY (K.J.G.); Department of Artificial Intelligence Research, Facebook, Montreal, Canada (M.D., A.R., M.R., P.V.); and Department of Computer Science, University of Florida, Gainesville, Fla (Z.Z.). Received February 28, 2019; revision requested April 9; revision received July 24; accepted August 29. Address correspondence to Y.W.L. (e-mail: Yvonne.Lui@nyulangone.org).

*F.K. and J.Z. contributed equally to this work.

In the last few years, there has been a substantial increase in research activity in the area of machine learning for MR image reconstruction (1–7), predominantly with the goal to accelerate MRI examinations by reducing the number of acquired k-space lines while still providing images with diagnostic quality or to enable imaging of dynamic processes with higher temporal resolution. These approaches train machine learning models with the goal of identifying the patterns of image artifacts that are introduced in the reconstructed images in accelerated acquisitions. The trained models are then used to reconstruct images from undersampled k-space data. However, the field has so far been constrained by the lack of a large-scale public dataset that includes raw k-space data. In the field of machine learning, large public datasets are routinely used for annual competitions and benchmarking (8). By contrast, MR image reconstruction studies are generally trained and validated on small isolated datasets compiled by independent groups and, in many cases, not shared with the greater research community. This has made it challenging to reproduce, validate, and meaningfully compare different approaches, and has limited the engagement of researchers outside of large centers where such data are available. The purpose of the fastMRI dataset is to provide the first step toward addressing this issue. Here we describe our recent release of the first large-scale dataset tailored to the problem of image reconstruction using machine learning techniques. Our dataset includes both raw MRI k-space data and magnitude Digital Imaging and Communications in Medicine (DICOM) images. The k-space data comprises 1594 measurement datasets obtained in knee MRI examinations from a range of MRI systems and clinical patient populations, with corresponding images derived from the k-space data using reference image reconstruction algorithms. The DICOM data represent an additional 10012 clinical image datasets from 9290 patients undergoing similar knee MRI examinations.

Description of the Dataset

The focus of our initial data release is to enable accelerated MRI acquisitions of two-dimensional (2D) fast-spin-echo sequences that are commonly used in musculoskeletal examinations. We include data from five sequences for different contrasts and image orientations that are used in the standard clinical knee examinations of our institution: (a) coronal proton density weighted, (b) coronal proton density weighted with fat suppression, (c) axial T2 weighted with fat suppression, (d) sagittal proton density weighted, and (e) sagittal T2 weighted with fat suppression.

The k-space dataset only contains the coronal acquisitions, and the range of sequence parameters are given in Table 1. The DICOM dataset contains data from all five sequences. Sequence parameters can be found directly in the DICOM headers of the data.

Curation of the dataset was part of a study approved by our local institutional review board. The k-space data were deidentified via conversion to the vendor-neutral International Society for Magnetic Resonance in Medicine (ISMRM) raw data format (9). DICOM data were deidentified by using the Radiological Society of North America's clinical trial processor tool (http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Processor). All metadata, as well as the DICOM images themselves, were manually inspected to ensure that no protected health information remained in the dataset.

The dataset is hosted in the cloud via Amazon web services and is available for download at https://fastmri.med.nyu.edu/. The total size of the k-space data is approximately 1.35 TB. It is split up into the following files for download: multicoil_train (931 GB), multicoil_val (192 GB), multicoil_test (109 GB), singlecoil_train (88 GB), singlecoil_val (19 GB), and singlecoil_test (7 GB). The total size of the combined DICOM image files is approximately 164 GB, and the files are stored with lossless JPEG 2000 image compression. The data are split up into the following files for

Abbreviations

DICOM = Digital Imaging and Communications in Medicine, ISMRM = International Society for Magnetic Resonance in Medicine, 2D = two dimensional

Summary

A publicly available dataset containing k-space and image data of knee examinations for accelerated MR image reconstruction using machine learning is presented.

Key Points

- The goal of this study was to share the fast MRI dataset to promote methodologic advances to enable large-scale validation of new algorithms and enhance reproducibility of scientific results in the field of MR image reconstruction.
- The fastMRI dataset contains both MRI k-space and DICOM (Digital Imaging and Communications in Medicine) image data obtained on knee MRI examinations.
- Increasing accessibility of MR images nationally and internationally can lead to the development of methods to reduce MRI scan time and image quality, which can both improve patient care and comfort.

download: DICOMs_batch1 (134 GB) and DICOMs_batch2 (30 GB). The dataset is hosted as tar.gz files, and the total size of these files is 4 GB smaller than the uncompressed files sizes. The dataset is open and available to anyone for educational and research purposes, with no requirement to submit a research proposal to access the data. However, users must sign up and agree to the data sharing agreement. The complete data sharing agreement is available on the download webpage. In general, while we do not allow any commercial use of the dataset itself, we do not explicitly discourage the development or testing of software, algorithms, or other intellectual property with the dataset.

k-Space Dataset

Fully sampled k-space data from 1594 consecutive clinical MRI proton density–weighted acquisitions of the knee in the coronal plane with and without frequency-selective fat saturation are included. The measurement identifiers in the k-space data were generated to be random integers. No examinations were excluded owing to presence of imaging artifacts from motion, pulsatile flow, and so forth. No contrast agents were injected at any of these examinations. Scans were performed on three clinical 3-T systems (Siemens Magnetom Skyra, Prisma, and Biograph-mMR) and one clinical 1.5-T system (Siemens Magnetom Aera) using clinical multichannel receive coils. Cartesian 2D turbo spin-echo sequences that are part of the routine clinical protocol at our institution were used. The publicly available software package Yarra (10) was used to gather k-space data from the MRI scanners. Example images from reference

reconstructions are shown in Figure 1. The data are provided together with metadata that allow reconstruction of images by means of a simple inverse Fourier transform. In particular, the individual k-space lines are already correctly sorted according to their position in the acquisition trajectory. No further preprocessing steps were performed on the data. An overview of the most relevant metadata fields is given in Table 2. For a complete list of the metadata that is included in the ISMRM raw data format, we refer the reader to Inati et al (9). The article that describes the k-space data format is also accompanied by an online code repository that provides tools to load and reconstruct the data for most commonly used programming languages and computing environments (C/C++, Matlab, Python). k-Space data from the fastMRI dataset can be processed with any of these code resources. Vendor-specific metadata about the pulse sequences used for data acquisition are not included. Because the data were acquired with a multichannel receive array coil, a proper combination of the individual coil images is a necessary step in the image reconstruction process. The most straightforward approach, which is also commonly used in clinical MRI protocols, is to use a sum-of-squares combination of the individual coil images. Image reconstruction of accelerated acquisitions via parallel imaging requires an additional calibration step to obtain coil sensitivity information. This can be done either explicitly by obtaining maps of the coil sensitivity profiles (11,12) or by estimating convolution kernels in k-space (13,14). Coil sensitivity profiles are not included in the database for two reasons. First, we want to avoid any bias toward a particular method for coil sensitivity estimation. The most common strategies for parallel imaging are described in the references cited above and several open-source software implementations for them are available online. Second, this would double the size of the dataset.

We also provide simulated single-coil k-space data derived from the acquired multicoil k-space data using an "emulated single-coil" combination algorithm (15). The rationale for providing simulated single-coil data—even though reconstruction

Table 1: Acquisition Parameters for the Imaging Protocols Used to Acquire Knee MRI Data Represented in the k-Space Dataset

Parameter	Coronal PD–weighted Protocol	Coronal PD–weighted FS Protocol
Sequence	2D turbo spin echo	2D turbo spin echo
FOV (mm²)	140 × 140	140 × 140
Matrix size	320 × 320	320 × 320
Turbo factor	4–5	4
Slice thickness (mm)	3	3
TR (msec)	2750–3000	2850–3000
TE (msec)	27–32	33
No. of slices	27–48	28–50
Receive coil	15 channel Tx-Rx	15 channel Tx-Rx

Note.—Because not all parameters are completely identical for the different MRI scanners that were used during data acquisition, a range of sequence parameters is shown in some cases. FOV = field of view, FS = fat saturation, PD = proton density, TE = echo time, TR = repetition time, Tx-Rx = transmit-receive, 2D = two dimensional.
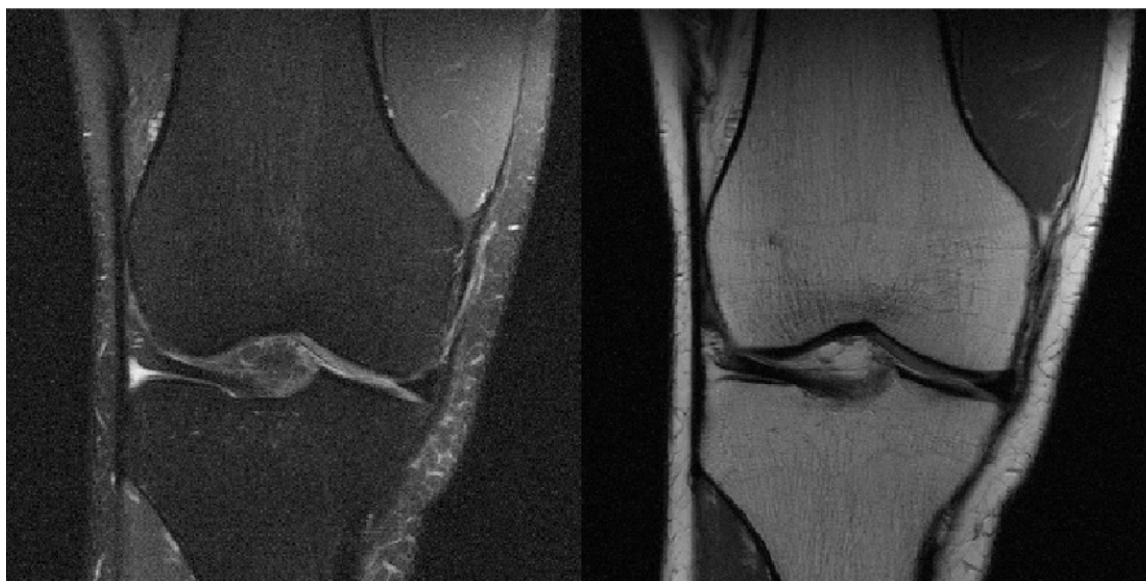
Figure 1: Coronal proton density–weighted images with fat suppression (left) and without fat suppression (right). Both images were reconstructed from fully sampled k-space data using a sum-of-squares combination of component coil images.

from multicoil data is expected to be more precise and closer to most clinical acquisition and reconstruction pipelines—is threefold: *(a)* To lower the barrier of entry for researchers who may not be familiar with MRI data, since the use of a single coil removes a layer of complexity, *(b)* to include a task that is relevant for the single-coil MRI machines still in use throughout the world, and *(c)* to separate out the aspects of reconstruction related to compressed sensing rather than parallel imaging. The 1594 k-space data examples are partitioned into the following six components: *(a)* training—coronal proton density weighted (484 examinations, average of 36 images) and coronal proton density weighted with fat suppression (489 examinations, average of 36 images); *(b)* validation—coronal proton density weighted (100 examinations, average of 36 images) and coronal proton density weighted with fat suppression (99 examinations, average of 36 images); *(c)* multicoil testing—coronal proton density weighted (59 examinations, average of 36 images) and coronal proton density weighted with fat suppression (59 examinations, average of 36 images); *(d)* single-coil testing—coronal proton density weighted (54 examinations, average of 36 images) and coronal proton density weighted with fat suppression (54 examinations, average of 36 images).

The remaining 196 examinations are held back for a planned image reconstruction challenge. The training and validation datasets may be used to fit model parameters and to optimize hyperparameter values. The test dataset is used to compare the results across different approaches. Evaluation on the test set is accomplished by uploading results to the public leaderboard at *https://fastmri.org/*. The first official challenge associated with the dataset is forthcoming.

The examples in the training and validation set are identical for the single-coil and multicoil datasets. For the challenge and test set, unique examples are provided for the single-coil and the multicoil dataset. This ensures that information cannot be shared between the two challenges at the test stage.

Table 2: Overview of Selected Metadata Fields That Are Included Together with the Raw k-Space Data	
Category	Raw Data Metadata Field (unit, if applicable)
Acquisition hardware	System vendor
	System model
	Field strength (T)
	Type of receive coil and number of elements
Encoded k-space	Matrix size (x,y,z)
	Field of view (x,y,z) (mm)
Reconstruction image space	Matrix size (x,y,z)
	Field of view (x,y,z) (mm)
Sequence parameters	Repetition time (msec)
	Echo time (msec)
	Inversion time (msec)
	Flip angle (degree)
	TSE echo spacing (msec)

Note.—A complete list of provided metadata in the International Society for Magnetic Resonance in Medicine raw data format is given in Inati et al (9). TSE = turbo spin echo.

Examples in the test and challenge sets contain undersampled k-space data. The undersampling is performed by retrospectively masking k-space lines from a fully sampled acquisition. k-Space lines are omitted only in the phase-encoding direction to simulate physically realizable accelerations in 2D data acquisitions. The same undersampling mask is applied to all slices of an example. To provide diverse undersampling patterns across the datasets, the undersampling mask is chosen randomly for each example, subject to constraints on the number of fully sampled central lines and the overall undersampling factor. Figure 2 shows details on the undersampling procedure.

DICOM Dataset

In addition to the k-space data, fastMRI also includes 10 012 consecutive DICOM image datasets from 9290 patients undergoing clinical knee MRI examinations with a full complement of clinical acquisitions represented, including a range of tissue contrasts and different planes of imaging. The total number of examinations and average number of images per sequence are: *(a)* coronal proton density weighted: 9947 examinations, average of 33 images; *(b)* coronal proton density weighted with fat suppression: 10 192 examinations, average of 31 images; *(c)* axial T2 weighted with fat suppression: 9640 examinations, average of 33 images; *(d)* sagittal proton density weighted: 10 491 examinations, average of 31 images; and *(e)* sagittal T2 weighted with fat suppression: 7311 examinations, average of 29 images.

This also includes contrast agent–enhanced examinations. Both pre- and postcontrast images are included. There is no overlap between the examinations in the DICOM dataset and the k-space dataset. No examinations were excluded owing to the presence of imaging artifacts from motion, pulsatile flow, and so forth. The DICOM data are produced using a wide range of scanners within our institution. They are not partitioned into training, validation, testing, and challenge sets. Instead, they are provided as a single dataset, for example, for the purpose of auxiliary training or to test generalizability of techniques developed. The DICOM patient identifiers were generated to be random integers.

Discussion

To our knowledge, this is the largest public dataset that includes raw k-space data and DICOM data from a clinical population. While public datasets of reconstructed images do exist—for example, the Human Connectome project (16), the Alzheimer's Disease Neuroimaging Initiative (17), and the Osteoarthritis Initiative (18)—they are generally specialized by already targeting a specific translational research question in which imaging serves as a tool to seek answers. The goal of our dataset is much broader: to provide a resource to improve image acquisition and reconstruction itself. Recent efforts have been devoted to collecting and publicly releasing datasets containing k-space data *(http://mridata.org* and *https://github.com/VLOGroup/mri-variationalnetwork).* However, the number of examinations that are provided in these datasets range between 10 and 100 and consequently might be too small for some machine learning–based reconstruction methods.
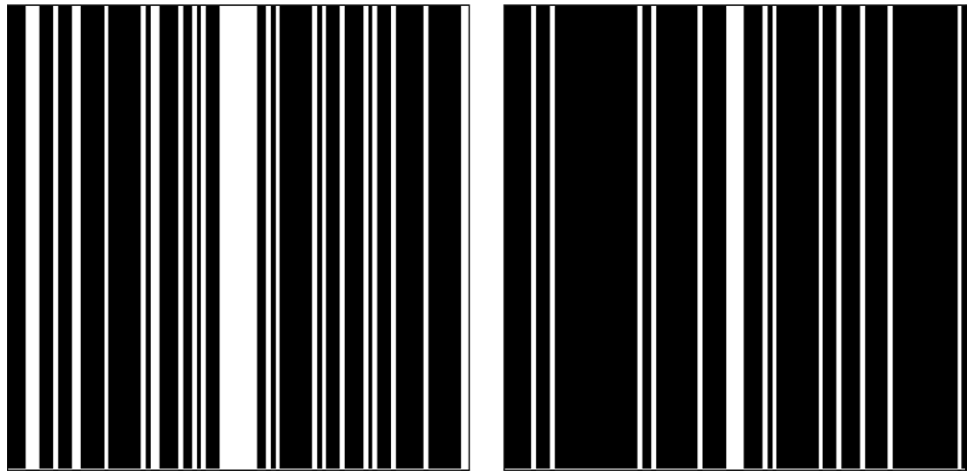

Figure 2: Examples of binary sampling masks (white = included, black = omitted) for pseudorandomly undersampled k-space data with fourfold acceleration (left) and eightfold acceleration (right). The overall acceleration factor is set randomly either to four or to eight (representing a fourfold or an eightfold acceleration, respectively), with equal probability for each example. The undersampling mask is then generated by first including some number of adjacent low-frequency k-space lines to provide a fully sampled central region of k-space. When the acceleration factor equals four, the fully sampled central region includes 8% of all k-space lines; when the acceleration factor equals eight, 4% of all k-space lines are included. The remaining k-space lines are included at random, by drawing samples from a uniform random distribution with the probability set such that the correct number of total k-space lines is achieved.

The fastMRI dataset is specialized at the moment because it is focused on 2D knee imaging data. We are planning to progressively add new data to the repository during future releases. Our next planned release will be for brain data and will follow an identical structure of both fully sampled k-space data and accompanying DICOM images. The dataset consists of consecutive examinations and therefore does include pathologic findings at a rate that is representative for a clinical patient population. However, because our focus in this project is on image reconstruction, we are currently not providing any diagnostic labeling segmentations, text reports, statistics on the prevalence of pathology, information on metal implants, or demographic information.

The number of cases included as DICOM images in our fastMRI dataset is substantially larger than the number of cases with k-space data. The DICOM portion of the dataset is also more heterogeneous, with data coming from a wider range of MRI systems and protocols. It is worth noting that a Fourier transform of these DICOM images does not directly correspond to the originally measured raw data. Many of the clinical images were acquired with accelerated acquisitions and reconstructed using approaches such as parallel imaging. In the context of machine learning for image reconstruction, our motivation to include the DICOM data is to answer the question of whether training on a larger number of imperfect examples can outperform training on a smaller number of high-quality examples. Complete details regarding this dataset, as well as relevant background material intended to empower investigators to tackle problems in image reconstruction, can be found in Zbontar et al (19).

We hope that the availability of this dataset can accelerate research in MR image reconstruction, much as the computer vision field was supercharged by well-curated large-scale natural image datasets such as ImageNet (8). In particular, we hope that

this dataset can serve as a benchmark for training and evaluation of new developments in image reconstruction, and that it can also serve as an example and a stimulus for the release of similar publicly available datasets in the future.

Author contributions: Guarantors of integrity of entire study, F.K., J.Z., M.B., M.P., Z.Z., D.K.S., Y.W.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, F.K., J.Z., A.S., M.J.M., K.J.G., Z.Z., M.D., A.R., M.R., P.V., E.O., M.P.R., D.K.S., Y.W.L.; experimental studies, F.K., J.Z., A.S., M.J.M., M.B., A.D., M.P., K.J.G., J.K., H.C., E.O., C.L.Z., M.P.R., Y.W.L.; statistical analysis, J.Z., M.J.M., J.P., C.L.Z.; and manuscript editing, F.K., J.Z., A.S., M.J.M., A.D., K.J.G., J.K., H.C., Z.Z., M.D., A.R., M.R., P.V., C.L.Z., M.P.R., D.K.S., Y.W.L.

References

1. Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. Magn Reson Med 2018;79(6):3055–3071.
2. Wang S, Su Z, Ying L, et al. Accelerating Magnetic Resonance Imaging Via Deep Learning. In: IEEE International Symposium on Biomedical Imaging (ISBI), 2016; 514–517.
3. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. Nature 2018;555(7697):487–492 .
4. Chen F, Taviani V, Malkiel I, et al. Variable-density single-shot fast spin-echo MRI with deep learning reconstruction by using variational networks. Radiology 2018;289(2):366–373.
5. Mardani M, Gong E, Cheng JY, et al. Deep generative adversarial neural networks for compressive sensing MRI. IEEE Trans Med Imaging 2019;38(1):167–179.
6. Schlemper J, Caballero J, Hajnal JV, Price AN, Rueckert D. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE Trans Med Imaging 2018;37(2):491–503.
7. Knoll F, Hammernik K, Kobler E, Pock T, Recht MP, Sodickson DK. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. Magn Reson Med 2019;81(1):116–128.
8. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009; 248–255.
9. Inati SJ, Naegele JD, Zwart NR, et al. ISMRM Raw data format: a proposed standard for MRI raw datasets. Magn Reson Med 2017;77(1):411–421.
10. Block TK, Sodickson DK. Yarra: An open software framework for clinical evaluation of reconstruction prototypes. In: ISMRM Workshop on Data Sampling & Image Reconstruction, 2016.
11. Pruessmann KP, Weiger M, Scheidegger MB, Boesiger P. SENSE: sensitivity encoding for fast MRI. Magn Reson Med 1999;42(5):952–962.
12. Pruessmann KP, Weiger M, Börnert P, Boesiger P. Advances in sensitivity encoding with arbitrary k-space trajectories. Magn Reson Med 2001;46(4):638–651.
13. Sodickson DK, Manning WJ. Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays. Magn Reson Med 1997;38(4):591–603.
14. Griswold MA, Blaimer M, Breuer F, Heidemann RM, Mueller M, Jakob PM. Parallel magnetic resonance imaging using the GRAPPA operator formalism. Magn Reson Med 2005;54(6):1553–1556.
15. Tygert M, Zbontar J. Simulating single-coil MRI from the responses of multiple coils. arXiv:1811.08839. [preprint] https://arxiv.org/abs/1811.08026. Posted 2018. Accessed January 2019.
16. Van Essen DC, Smith SM, Barch DM, et al. The WU-Minn Human Connectome Project: an overview. Neuroimage 2013;80:62–79.
17. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology 2010;74(3):201–209.
18. Eckstein F, Wirth W, Nevitt MC. Recent advances in osteoarthritis imaging--the osteoarthritis initiative. Nat Rev Rheumatol 2012;8(10):622–630.
19. Zbontar J, Knoll F, Sriram A, et al. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. arXiv 1811.08839. [preprint] https://arxiv.org/abs/1811.08839. Posted November 2018. Revised December 2019.