Journal of Nuclear Medicine, published on November 22, 2016 as doi:10.2967/jnumed.116.177303

Small Data: a ubiquitous, yet untapped, resource for low cost imaging innovation

Adam Leon Kesner, PhD[1], Wolfgang Weber, MD[2]

[1] Department of Radiology, Division of Nuclear Medicine and Molecular Imaging, University of Colorado School of Medicine, Aurora, CO, USA

[2] Department of Molecular Imaging and Therapy, Memorial Sloan Kettering Cancer Center, New York, New York, USA

Corresponding author:
Adam Kesner, Ph.D.,
Medical Physicist and Assistant Professor (Specialty Nuclear Medicine)
Department of Radiology, Division of Radiological Sciences
University of Colorado Denver, School of Medicine, Anschutz Medical Campus
12700 E 19th Ave, Box C-278
Aurora, CO 80045
Office: (303) 724-3764
Fax: (303)-724-3795
Email: adam.kesner@ucdenver.edu

Short running title: Small data innovation

Key words: Small Data, Data driven innovation, Medical Imaging, PET

Word count (text + references): 2100 words

Positron Emission Tomography (PET), 'conventional' nuclear imaging, and the vast majority of contemporary medical imaging modalities, are inherently digital technologies. Over the last several decades there has been a transformative evolution of the digital computing landscape with respect to speed, cost of storage, infrastructure and expertise. However our use of data, and in fact our whole understanding of the role of data in relation to emission imaging has remained relatively unchanged. If we take a moment to reflect on this resource, generated ubiquitously in our daily imaging procedures, we can recognize that we have the capacity to support information utilization beyond present convention and that the 'raw' data of nuclear imaging studies provide can be tapped to fuel innovation.

Our general understanding of image data is that it exists in the form of Digital Imaging and Communications in Medicine (DICOM) format images, essentially analogous to film and representing a quantity of source signal distributed in space. However, the signals and information used to create these images in nuclear medicine originate in a much denser form – our imaging machines capture highly detailed time, location, and energy information for individual decay events. The current practice in PET for example, is to truncate this information using assumptions and reconstruction techniques, so as to provide a representation of tracer emissions distributed in recognizable Cartesian space. This process of biodistribution-representative image generation has essentially defined nuclear imaging for half a century. The procedure of truncating (unused) information is heavily ingrained in our practice likely due to the fact that for most of the field's existence it has been expensive and impractical to save raw acquisition data.

The costs associated with saving data have never been a static consideration. In 1980, a Gb of data cost $600,000 (*1*) (approximate value, inflation adjusted), in 1990 that cost went down to $15,000, in 2016 it costs $0.02, and we can confidently project continuation of this financial trend. Retaining a 2 Gb raw PET acquisition file now represents approximately 0.001% of the market cost of a scan. Both the cost and capacity of digital imaging has undergone a slow but in aggregate very large shift. Each year data driven

solutions become more practical and more relevant than the last, as shown in Figure 1. In the 1990s we passed a milestone when digital storage became more cost effective than paper (*2*). It is possible that we have now passed a new barrier in that we can say the cost of saving raw imaging acquisition data is negligible relative to the cost of generating it. Furthermore, with ionizing radiation imaging the cost of data paradigm is not only economic – patients are being exposed to radiation to generate this data and at risk to their health, it is prudent for us to periodically reconsider if our practices are making optimal use of it.

One reason to support changing our data saving practice towards more robust access and archiving comes from the fact that we already have a body of literature showing that access to raw data can enable creative innovation. As an example, our group has recently published a study showing that large populations of scans can be corrected for motion using advanced data utilization techniques and without the need for gating equipment or modified acquisition procedures (*3*). Additional areas of respiratory, cardiac, and head motion correction, signal optimization, dose optimization, open source reconstruction, and retrospective reframing have also begun to be explored (*4*). Progress in these areas and the impact of data based innovation efforts have been limited because of the rigid data access framework we currently have in place. We delineate acquisition data as proprietary, which subsequently impedes the academic and commercial exploration that traditionally propels potentially impactful ideas beyond the labs they are created in. For example commercial or open source standardized PET reconstruction techniques could enable new levels of image quantification standardization across PET machines/centers and boost statistical power in research and multi-center clinical trials (*5*). Standardization can also make a large impact in the newly emerging fields of big data (*6*), machine learning (*7*), and radiomics (*8*) – arguably there is no single effort that would more positively affect the latter two areas of research. Looking in another direction, we can also recognize personalized image reconstruction as a promising area of study, with efforts for task based optimization emerging (*9*) and

demonstrating the importance of researcher access to raw data. With respect to industrial innovation, accessible raw data would mean that those wishing to innovate in its use commercially, for example to build a distributable data driven gating (*3,10*) solution, would have access to raw data to develop the technologies, incentive to obtain regulatory approvals, and the ability to distribute solutions throughout the community. The third party DICOM-based innovations we have seen developed provide an example of how access to data in combination with commercial/market forces can bring creative solutions into clinical use.

Another argument to support changing our data saving practices is that our field benefits when it cultivates low cost innovation. Lower costs generally mean greater inclusion of the research community for developing and benchmarking products and ultimately greater access for clinical end users. Looking forward there is also an imperative to ensure our field continues to provide relevant leadership in future. Attention to both costs and benefits of our innovations is important for ensuring our technologies provide relevant benefit in clinical care. In the United States and globally we are witnessing efforts to combat the inflating costs of healthcare. Traditional strategies of investing in expensive powerful technologies and waiting for cost-effectiveness to catch up may not be as viable a path for innovation as it has been in years past. Data driven innovation stands in contrast to a hardware driven model, as it is centered on a concept of "doing more with what you (already) have". What's more, the trend of digital evolution not only empowers us to develop innovative uses of data, but we can also be assured that if we tie our solutions to data utilization technologies we can reasonably expect them to get faster, cheaper and more powerful with time. Finally, nuclear medicine is a global field, and recognizing that data produced in all our systems is a valuable tool for innovation would enable inclusive paths toward innovation that expand our pool of creative talent and potential leaders.

The final rationale we present to support changes in our data saving practices addresses what we don't know about the future. Even if there were insufficient argument for changing our practice today, we can

look forward and see how changes now may benefit the pioneers of tomorrow. What if for example, we were better stewards of medical data in the 1990s – we would now be decades ahead in our efforts to mine and interpret big data. We are now changing our practices to archive electronic medical records and image data, but we are not commonly archiving raw data. We admittedly don't know what the future will bring. However, recognizing now the potential value of raw data, identifying it as a resource, and preparing it to be harvested by future generations is within our present capacity. It is not difficult to imagine how very near-future innovations in radiomics and computer aided diagnosis would benefit greatly from access to raw data along with archived medical records for benchmarking new techniques in large, standardized, associative studies. Looking forward we can also consider that the newly trained and future generations of imaging professionals, as well as the patients they are serving, will be 'digital natives', and will likely have talents and expectations for data management that go beyond the current standard.

The benefit of data valuation extends across the medical imaging fields. However, the nuclear medicine community is favorably positioned to play a leading role in redefining the value of raw acquisition image data. Our data is inherently filled with useful timing, energy, and spatial information; nuclear imaging is used for a variety of applications that span the medical specialties; our field has long included cooperation among a variety of specialists within our community - physicians, physicists, computer scientists, mathematicians, and others. This diversity and history of cross-specialty collaboration places our field in a favorable position to pioneer new concepts on the value of data and advanced data utilization based innovation.

In an effort to coalesce the ideas aforementioned in this letter, we to take this opportunity to present the concept of *small data* for the nuclear medicine/imaging communities:

Small data is defined as informative, possibly ancillary details inherent within data or data sets. Small data is local, actionable, often personalized elements of information that can inform and enable optimal utility of its instance. The term *small data* encompasses the notion that every digital bit of information may have value and utility, and implicitly implies the importance of its access.

Currently there is enormous excitement about using the enormously increased power of computer systems to mine 'big data', i.e. data that is stored across multiple databases and captures information from various sources, such as genetics, treatments and long-term patient outcome. We argue that modern computing power should also be applied to 'small data', the raw data that are routinely acquired every day during clinical practice.

The term *small data* in not entirely new, at present various mentions can be found on the Internet. To our knowledge however, it has not been defined as a concept of data valuation formally or in peer reviewed publications. We are taking this opportunity to do that, and to clarify its relevance to the imaging community. Small data innovation encompasses technologies that utilize small data details, and represents an area with potential for meaningful imaging innovation. Small data innovation may include revisiting traditional uses of data and extracting greater details and/or using modern computing power to develop improved processing strategies. Small data innovation can support personalized imaging or task optimized imaging through advanced information extraction techniques, personalized image processing and dose optimization.

In the 20th century a main challenge of nuclear medicine and radiology IT was reconstructing 3D-images, and displaying and storing the reconstructed images efficiently. As we progress into the 21st century, the vastly increasing processing power, network bandwidth and storage capacity of current computer systems now allow us to go beyond the reconstruction of the distribution of radioactivity at a given period of time. By storing and analyzing the raw images we can derive additional information, such as

signal from motion, spatial/temporal characterization of signal dependability, or quantitative uptake measurements based on open source standardized reconstruction strategies. By reviewing and updating data access practices we can open the door to new, clinically applicable commercial innovations – in the same manner in which we found success with the DICOM image standardization initiative of the 1990s. What is required now is not expensive investment, but rather an open mind in our community towards exploring the potential benefits of understanding and using 'small data' differently.

References:

1. Cost of hard drive storage space. http://ns1758.ca/winch/winchest.html. Accessed 8/15/2016, 2016.

2. Morris RJT, Truskowski BJ. The evolution of storage systems. *IBM Systems Journal.* 2003;42:205-217.

3. Kesner AL, Chung JH, Lind KE, et al. Validation of software gating: a practical technology for respiratory motion correction in PET. *Radiology.* 2016:152105.

4. Kesner AL, Daou D, Schindler TH, Koo PJ. Carpe datum: a consideration of the barriers and potential of data-driven PET innovation. *J Am Coll Radiol.* 2016;13:106-108.

5. Clarke LP, Nordstrom RJ, Zhang H, et al. The Quantitative Imaging Network: NCI's historical perspective and planned goals. *Transl Oncol.* 2014;7:1-4.

6. Kansagra AP, Yu JP, Chatterjee AR, et al. Big data and the future of radiology informatics. *Acad Radiol.* 2016;23:30-42.

7. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal.* 2012;16:933-951.

8. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* 2016;278:563-577.

9. Armato IIISG, Hadjiiski L, Tourassi GD, et al. Guest Editorial: LUNGx Challenge for computerized lung nodule classification: reflections and lessons learned. *Journal of Medical Imaging.* 2015;2:020103-020103.

10. Kesner AL, Bundschuh RA, Detorie NC, Dahlbom M, Czernin J, Silverman DHS. Respiratory gated PET derived from raw PET data. Paper presented at: Nuclear Science Symposium Conference Record, 2007. NSS '07. IEEE; Oct. 26 2007-Nov. 3 2007, 2007.

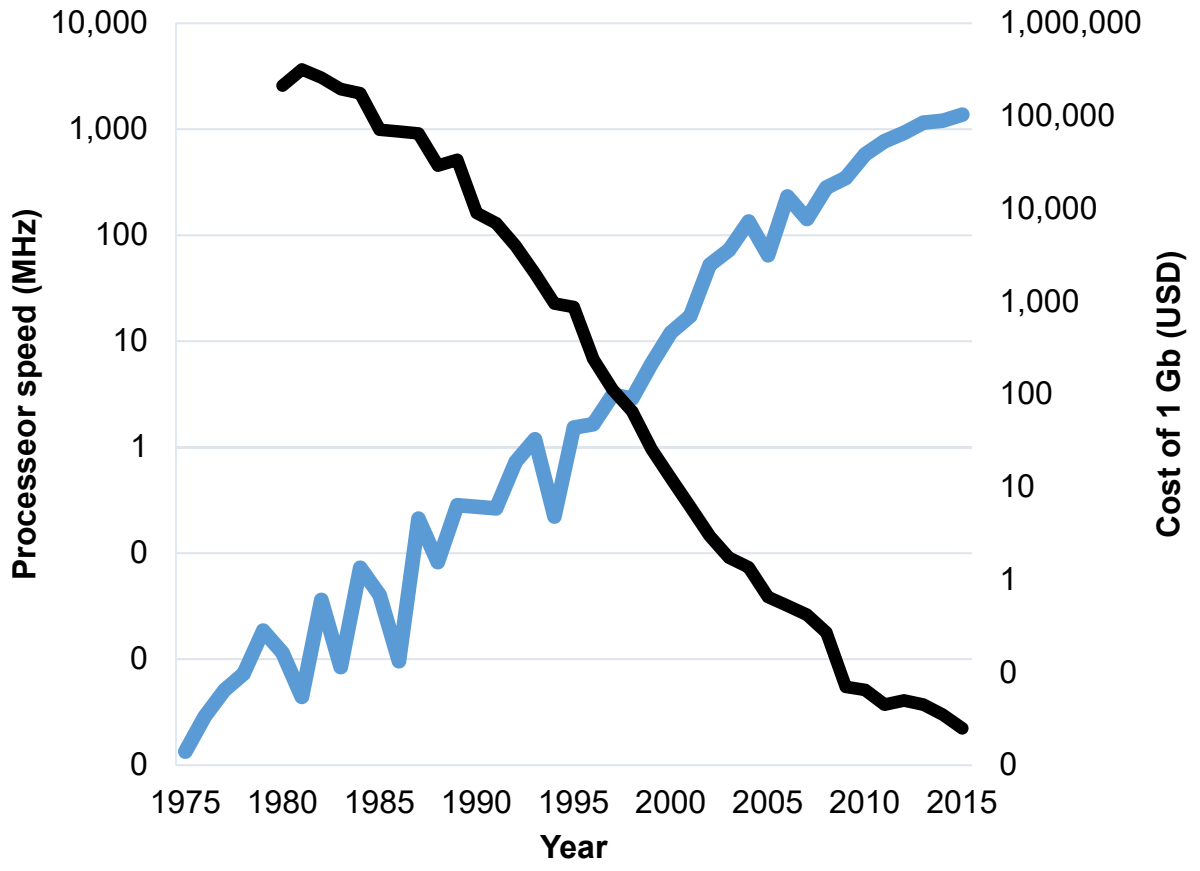

Figure 1- PET is supported by a digital infrastructure that has undergone large transformations in the last decades. Illustrated here are the speed of processors (blue) and cost of storage (black) (*1*) shown sample averaged across years. All data shown on log scales. (processor speed extrapolated from collection of historical transistor count references hosted on Wikipedia page - https://en.wikipedia.org/w/index.php?title=Transistor_count&oldid=734427585, accessed 10/5/2016)