The Whole Patient

Big data for deep learning in radiology: population-based imaging

How artificial intelligence is about to transform radiological research and clinical care

Big data and deep learning techniques are increasingly being applied to radiological problems, where they serve as powerful complements to traditional study designs.

Christopher Schuppert

Radiology Resident
Department of Diagnostic and
Interventional Radiology
University Hospital Heidelberg
Heidelberg, Germany

christopher.schuppert@
med.uni-heidelberg.de

Fabian Bamberg

Professor and Chairman
Department of Diagnostic and
Interventional Radiology
University Hospital Freiburg
Freiburg, Germany

fabian.bamberg@
uniklinik-freiburg.de

Hans-Ulrich Kauczor

Professor and Chairman
Department of Diagnostic and
Interventional Radiology
University Hospital Heidelberg
Heidelberg, Germany

hans-ulrich.kauczor@med.
uni-heidelberg.de

Christopher L. Schlett

Division Chief for
Cardiovascular Imaging
Head of Population-based Imaging
Department of Diagnostic and
Interventional Radiology
University Hospital Heidelberg
Heidelberg, Germany

christopher.schlett@med.
uni-heidelberg.de

The concept of artificial intelligence (AI) has been around for more than 50 years and has recently found its way into medicine, particularly into the field of radiology. It is currently widely discussed in both scientific and non-scientific communities as a potential game changer likely to improve diagnostic accuracy, efficiency and effectiveness along with paving the way towards personalised medicine.

Particular focuses of these discussions are machine learning and its sub-discipline deep learning, which is one of the latest developments in the constantly evolving field of AI. In deep learning models, multiple layers of artificial neural networks are linked to create complex algorithms most commonly in the form of "convolutional neural networks". The exponential increase in the amount of data required to train these systems is leading scientists and companies to seek out new sources of sufficiently large data sets and is serving as a major driver of big data.

Big data and the 'black box' problem

Although there is no formal definition of big data, a description proposed by the McKinsey Global Institute is widely agreed upon. It characterises big data as "data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse" (Manyika et al. 2011). In the context of big data, the 5 Vs should be considered: (1) Volume, which quantifies the number of records and/or storage size; (2) Variety, which specifies the number of sources data is drawn from; (3) Velocity, which measures the frequency of new data being added; (4) Veracity, which describes the quality and accuracy of the acquired data; and (5) Value, which assesses the economic benefit of the inferred conclusions (White 2012; Kwon and Sim 2013; Fosso Wamba et al. 2015). Several of these features have the potential to introduce noise and errors in databases, including those of medical images. Although a well-adjusted algorithm has the ability to overcome such variations in the data much better than traditional prediction models, the initial training of its underlying deep neural network is highly sensitive to errors, and the algorithm may consequently be prone to misclassifications. Moreover, trained neural networks have a "black box" problem: while the prediction matrix of a traditional model may well be comprehended against its input characteristics, the computations performed by deep neural networks can be so arbitrarily complex that there is no simple link between the weights of the input characteristics and the function being approximated.

Thus, when talking about big data for deep learning in radiology, we need to particularly aim for changes affecting two Vs—yielding increased veracity and decreased variety. There are several sources of medical images, which if separated by origin, may be stratified into clinical data and research data. Clinical imaging data is abundant [it is estimated that digital healthcare data will amount to 2,314 exabytes (2.3 billion gigabytes) globally by 2020 (IDC 2014), with the vast majority coming from clinical images] and reflects real-world problems. However, it suffers from both high variety and low veracity. Promising alternative data sources are large-scale, population-based cohort studies that employ standardised protocols for medical imaging.
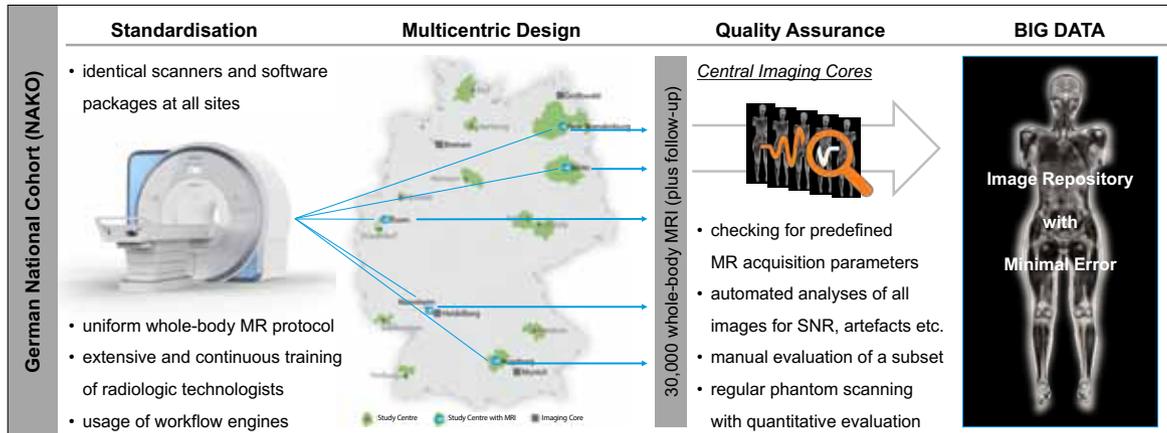
Figure 1. Population-based imaging with a high level of standardisation and quality assurance as a source for big data, exemplified by the German National Cohort (NAKO).

NAKO is a population-based cohort study including 18 sites across Germany; a subset of about 30,000 participants will undergo whole-body MR imaging at 5 different sites. After baseline, follow-up examinations will be conducted including repeated MR scans after 4-5 years. Large extent of standardisation and intensified quality assurance by dedicated imaging cores guarantee minimal random error in the image repository—a significant source of big data for deep learning. SNR denotes signal-to-noise ratio.

> ❝ THE STRUCTURED DESIGN OF SUCH POPULATION-BASED IMAGING STUDIES ALLOWS FOR A DECREASE IN VARIETY AND AN INCREASE IN VERACITY ❞

Structured population-based imaging studies

Population-based imaging in large cohort studies is increasingly being used in epidemiological studies with the aim of identifying physiologic variants and detecting subclinical disease burden. Several such studies are ongoing. Among the most prominent examples are two European projects: the UK Biobank (aiming for 100,000 whole-body MR scans among 500,000 total participants) and the German National Cohort (NAKO; aiming for 30,000 whole-body MR scans among 200,000 total participants). The NAKO study is divided into 18 study sites across Germany and the MR substudy is performed at 5 of those locations, where identical 3.0 Tesla scanners have been installed (**Figure 1**). The uniform imaging protocol used at each site comprises a set of 12 native series covering the neurological, cardiovascular, thoracoabdominal and musculoskeletal body regions with a total scan time of approximately 1 hour (Bamberg et al. 2015). In addition to the cross-sectional baseline study that began in 2014 and is projected to end in 2018, follow-up examinations including repeated MR scans will be conducted after 4-5 years and will permit supplementary longitudinal analyses. Embedding the MR substudy into the larger main study allows for imaging data to be correlated with a series of other assessments, including eg neurocognitive function tests, blood pressure measurements, electrocardiography, echocardiography and a range of biomaterials, such as blood samples or DNA. Drawing from the extensive data repository outlined above, these projects are, among other insights, expected to make valuable contributions to the field of radiomics by identifying and validating new imaging biomarkers with a predictive value for future clinical disease manifestation.

The structured design of such population-based imaging studies allows for a decrease in variety and an increase in veracity. With respect to the NAKO Cohort, the identical MR scanners and their unified imaging protocol help to eliminate technical variances almost entirely. Centralised staff training and certification, in conjunction with real-time, automated quality assurance for completeness of data, conformity of scan regions and scan parameters as well as adequateness of global image features (such as signal-to-noise ratio and a mathematically defined universal image quality index) effectively warrant high veracity of the data (**Figure 1**). The cross-sectional study approach means forgoing real-time data and limits data volume to a predefined amount, but implicitly keeps data velocity at a controllable pace and data variety at a manageable level. Although these arrangements contradict traditional big data attributes to some extent (the above-mentioned 5 Vs), they enable consistency and will thereby facilitate research in a field where several previous big data approaches have failed, particularly due to data complexity and high level of error.

The application of big data techniques is a relatively novel endeavour in radiology compared to other industries and fields of research. The objective to build up large imaging biobanks and to extract previously unidentified features somewhat replicates the ventures undertaken several years ago in the distantly related areas of genomics and proteomics, but comes with a unique set of challenges. MR imaging has already emerged as the modality of choice for dedicated population-based cohort studies, yet other aspects such as the accomplishment of reliable and accurate organ segmentations remain challenging and require further consideration (Schlett et al. 2016). Novel segmentation algorithms based on deep learning approaches are currently being developed for and on the standardised acquired MR data from population-based cohort studies. These algorithms will certainly also have a direct impact on the clinical work of radiologists, as will the comprehensive imaging data itself: with whole-body MR scans being performed in an increasing number of population-based cohort studies, we will be able to not only inspect isolated organs but also to develop novel knowledge on how diseases affect multi-organ systems and/or whether subclinical alterations of one organ are a risk marker for the impairment of others. Consequently, multi-level relationships of diseases can be studied and image-based risk stratification systems suitable for personalised medicine may be established. The insights gained will benefit research as well as clinical care. Another matter of discussion touches on the subject of evidence-based medicine, which also encompasses evidence-based radiology: we need to ask ourselves whether or not the previously mentioned black box problem is a tolerable issue for determining pathophysiological links based on correlations derived from deep learning algorithms.

Conclusion

As with any pioneering work, there are many more topics that are still up for debate. In a larger sense, however, the application of big data and deep learning methods in radiology is holding a vast and as yet largely untapped potential for significant advancements in healthcare. Traditional study designs along with their methods of data collection and analysis remain important and we aim to promote the synergistic effects between these two scientific paradigms—capable of serving as powerful supplements to, rather than substitutes for one another. Large population-based cohort studies involving whole-body MR imaging, such as the NAKO, embrace this synergy and may function as a fruitful source of big data for deep learning in radiology. ■

Note:

The authors' institutions both take part in the NAKO study. FB and CLS serve as head of the MRI substudy.

KEY POINTS

✓ Within the healthcare sector, radiology is taking a leading role in implementing artificial intelligence solutions into research and clinical imaging

✓ Big data concepts rely on the "5 Vs" as key variables: Volume, Variety, Velocity, Veracity, and Value

✓ The German National Cohort (NAKO) features an embedded MR substudy that was designed in accordance with big data and deep learning requirements

✓ Large-scale population-based imaging research can benefit heavily from big data concepts and deep learning methods

REFERENCES

Bamberg F et al.; German National Cohort MRI Study Investigators (2015) Whole-Body MR imaging in the German National Cohort: rationale, design, and technical background. Radiology 277 (1): 206-20.

Fosso Wamba S et al. (2015) How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics 165: 234-46.

EMC (2014) EMC digital universe infobrief with research and analysis by IDC. The digital universe of opportunities: rich data and the increasing value of the internet of things. Available from emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

Kwon O, Sim JM (2013) Effects of data set features on the performances of classification algorithms. Expert Systems with Applications 40 (5): 1847-57.

Manyika J et al. (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey. Available from mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Schlett CL et al. (2016) Population-based imaging and radiomics: rationale and perspective of the German National Cohort MRI Study. Rofo 188 (7): 652-61.

White M (2012) Digital workplaces: vision and reality. Business Information Review 29 (4): 205-14.