European Radiology
Experimental

ORIGINAL ARTICLE   **Open Access**

CrossMark

Generative models: an upcoming innovation in musculoskeletal radiology? A preliminary test in spine imaging

Fabio Galbusera[1][*] , Tito Bassani[1], Gloria Casaroli[1], Salvatore Gitto[2], Edoardo Zanchetta[2], Francesco Costa[3] and Luca Maria Sconfienza[2,4]

Abstract

Background: Deep learning is a ground-breaking technology that is revolutionising many research and industrial fields. Generative models are recently gaining interest. Here, we investigate their potential, namely conditional generative adversarial networks, in the field of magnetic resonance imaging (MRI) of the spine, by performing clinically relevant benchmark cases.

Methods: First, the enhancement of the resolution of T2-weighted (T2W) images (super-resolution) was tested. Then, automated image-to-image translation was tested in the following tasks: (1) from T1-weighted to T2W images of the lumbar spine and (2) vice versa; (3) from T2W to short time inversion-recovery (STIR) images; (4) from T2W to turbo inversion recovery magnitude (TIRM) images; (5) from sagittal standing x-ray projections to T2W images. Clinical and quantitative assessments of the outputs by means of image quality metrics were performed. The training of the models was performed on MRI and x-ray images from 989 patients.

Results: The performance of the models was generally positive and promising, but with several limitations. The number of disc protrusions or herniations showed good concordance ($\kappa = 0.691$) between native and super-resolution images. Moderate-to-excellent concordance was found when translating T2W to STIR and TIRM images ($\kappa \geq 0.842$ regarding disc degeneration), while the agreement was poor when translating x-ray to T2W images.

Conclusions: Conditional generative adversarial networks are able to generate perceptually convincing synthetic images of the spine in super-resolution and image-to-image translation tasks. Taking into account the limitations of the study, deep learning-based generative methods showed the potential to be an upcoming innovation in musculoskeletal radiology.

Keywords: Lumbar vertebrae, Machine learning (deep learning), Magnetic resonance imaging, Neural network (computer), X-rays

Key points

- Deep learning-based generative models are able to generate convincing synthetic images of the spine
- Generative models provide a promising improvement of the level of detail in MRI images of the spine, with limitations requiring further research
- The availability of large radiological datasets is a key factor in improving the performance of deep learning models

Background

Artificial intelligence and deep learning are ground-breaking technologies that are revolutionising several research and industrial fields. The most notable current uses of deep learning include tasks such as computer vision, pattern recognition of images and classification of complex data, which are increasingly commonly used in sophisticated applications such as robotics, self-driving

* Correspondence: fabio.galbusera@grupposandonato.it
[1]Laboratory of Biological Structures Mechanics, IRCCS Istituto Ortopedico Galeazzi, via Galeazzi 4, 20161 Milan, Italy
Full list of author information is available at the end of the article

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 2 of 13

cars and automated computer programming [1, 2]. Among deep learning technologies, generative models, i.e. computer programs able to generate novel data rather than classify or process existing data, are recently gaining interest due to major technical innovations which are dramatically improving their performance [3–5]. The use of conditional generative adversarial networks (GANs) has recently become a standard for supervised image-to-image translation, i.e. the automatic conversion between two types of images [4]. For example, conditional GANs have been used to colourise black and white pictures and to generate photorealistic images from schematic ones and vice versa [6]. A decisive factor for the current popularity of such technologies is the free availability of most of these computer codes, which are usually written in the popular Python language, as well as of the underlying computational libraries, such as Tensorflow (https://www.tensorflow.org/), Caffe (http://caffe.berkeleyvision.org/), Torch (http://torch.ch/) and PyTorch (https://pytorch.org/). Besides, the training of small- and medium-scale deep learning models does not require specialised hardware, as it can be performed on a standard personal computer equipped with a modern graphics processing unit.

Nowadays, the application of deep learning techniques and generative models in musculoskeletal radiology is still in its infancy. Regarding spine imaging, several studies described the use of simpler machine learning methods such as multilayer perceptron and support vector machines for regression and classification problems, such as the identification of landmarks and the grading of intervertebral disc degeneration [7–9]. More clinically oriented studies—for example, those aimed at the prediction of postoperative complications and patient satisfaction or at the spinal alignment after spine deformity correction surgery—have also been published [10–12]. However, the most recent deep learning techniques, including generative models, did not significantly impact the methods and procedures employed in musculoskeletal radiology so far.

We recently explored the use of conditional GANs for the generation of synthetic radiological images of the spine, namely coronal and sagittal x-rays to be used in in silico trials, i.e. simulated trials aimed to predict the clinical performance of novel implants and surgical techniques before the final validation in a real clinical trial [13]. Nevertheless, generative models likely have the potential to be also employed in the standard diagnostic imaging, by improving the quality of the outputs of the imaging systems as well as by providing additional information which is not accessible in the original images. The aim of the present study was, therefore, to test the potential of generative models in diagnostic musculoskeletal imaging by performing a small set of clinically relevant benchmark cases in the field of spine imaging.

Methods

Ethical committee approval for this retrospective study has been obtained and patients' informed consent was waived. All magnetic resonance imaging (MRI) examinations were performed using one of two 1.5-T systems installed at our institution (Avanto, gradient strength 45 mT/m, slew rate 200 T/m/ms; or Espree, gradient strength 33 mT/m, slew rate 170 T/m/ms; Siemens Medical Solutions, Erlangen, Germany).

Two distinct potential applications of conditional GANs in the field of spinal imaging were tested in the present work. First, the enhancement of the resolution of T2-weighted (T2W) images (super-resolution) was performed. Then, the automated translation between different imaging modalities, referred to in the following paragraphs as virtual multimodal imaging, was attempted. The sizes of the training, validation and test datasets for each task are summarised in Table 1.

Conditional GANs and Pix2pix

Standard, i.e. unconditional, GANs consist of two components: a model, named the discriminator, which is trained to discriminate whether an image is real or fake; and a generator, which learns to create realistic new data trying to fool the discriminator, with the final aim of 'making the output indistinguishable from reality' [4, 5]. In other words, the discriminator teaches the generator how to create a realistic output, while simultaneously learning how to discriminate real and fake data. Unconditional GANs have been used for image-to-image translation problems by designing a loss function, such as L1 or L2 regression or more complicated application-specific functions, to link the output of the GANs to the input [14, 15]. Regression-based loss functions consider each pixel in the output as independent, not accounting for its relation to the surrounding parts of the image, and thus favours the optimisation on a pixel-to-pixel basis rather than of the image structure as a whole.

Whereas in unconditional GANs the generator does not see the input image, in conditional GANs both generator and discriminator receive as input the image to be translated [4, 16]. The loss function of conditional GANs is

Table 1 Sizes of training, validation, and test datasets for all the considered tasks

Task	Training	Validation	Test
Super-resolution	767	192	30
T1-weighted to T2-weighted	767	192	30
T2-weighted to T1-weighted	767	192	30
T2-weighted to STIR	284	71	30
T2-weighted to TIRM	305	77	30
Sagittal x-ray to T2-weighted	363	91	30

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 3 of 13

learned, rather than predefined. Research showed that these architectural improvements provided clearly superior performances with respect to GANs using only traditional regression loss functions [4], which tend to produce blurry outputs [17]. Indeed, conditional GANs now constitute the state-of-the-art for image-to-image translation.

In this work, we used pix2pix [4], the original formulation of conditional GANs for image-to-image translation, in its Tensorflow implementation (available at https://github.com/affinelayer/pix2pix-tensorflow), which closely follows the original implementation published by the authors based on Torch (https://github.com/phillipi/pix2pix). This port does not include the 'U-Net' architecture with skips for the generator [18] described in the original paper [4], and allows for combining the conditional GAN objective function with a L1 regression loss, as desired by the user. As suggested by Isola et al., we used a combination of loss weights (1 for the conditional GAN loss, 100 for the L1 loss) which were shown to ensure a sharp output while limiting the occurrence of visual artifacts. Since the original implementation only supports images with a size of 256×256 pixels, for the task involving images of 512×512 pixels we enlarged the encoder–decoder networks of both the generator and the discriminator to allow the processing of such images. All models were trained from scratch on a Linux workstation equipped with a NVIDIA Titan Xp GPU. Source codes and pre-trained models used and developed in this study are publicly available at https://goo.gl/xAgkbd.

Super-resolution
From a large set of high quality T2W images of the lumbar spine available in the picture archive and communication system (PACS) of IRCCS Istituto Ortopedico Galeazzi, 989 midsagittal images with a size of 512×512 pixels were collected. All images were linearly down-sampled to a size of 128×128 pixels and then linearly re-upscaled to 512×512 pixels to obtain a low-resolution image containing no fine details. Conditional GANs were then trained to reconstruct a sharp image of 512×512 pixels, starting from the corresponding low-resolution one.

Virtual multimodal imaging
The same technique used for super-resolution, i.e. the standard image-to-image translation by means of conditional GANs, was tested in the automated translation between different imaging modalities. Five tasks were attempted with this technique: (1) the translation from T1-weighted (T1W) to T2W images of the lumbar spine; (2) the reverse translation, i.e. from T2W to T1W images; (3) from T2W to short tau inversion recovery (STIR) images; (4) from T2W to turbo inversion recovery magnitude (TIRM) images; (5) from sagittal standing x-ray projections to T2W images. For the first four tasks, the

training datasets were created based on a database of 989 patients subjected to T1W and T2W imaging of the lumbar spine for the investigation of low-back pain available in our PACS. In the image database, STIR imaging was available for 385 individuals, while TIRM images were acquired for 412 patients. All images had a size of 256×256 pixels. To maximise the size of the training datasets, in addition to the midsagittal images of the lumbar spine, the adjacent images were also included, for a total of three images for each patient.

For the fifth translation task aimed at the generation of T2W images from sagittal standing x-ray projections, 484 patients for whom images acquired in both modalities were available in our PACS were identified. After resizing all images to 256×256 pixels, all planar radiographic projections were registered to the corresponding midsagittal MR image, so that the corners of the vertebral body of L5 were approximately in the same location in both images (Fig. 1). To this aim, an in-house C++ computer program which allows for the manual identification of the corner points of L5 and then performs a Euclidean transformation of the images was developed and used to align each couple of images for the 484 patients. Subsequently, a generative model able to translate the radiographs to the registered T2W midsagittal MR images was trained.

Quantitative validation
All image-to-image translation models were tested on 30 images for which the ground truth, called 'target', was available. The quality of the generated outputs was assessed by means of commonly employed metrics such as the mean squared error (MSE), the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM) [19] and the visual information fidelity in pixel domain (VIPF) [20], taking as reference the target image. Furthermore, the cumulative probability of blur detection (CPBD) [21], which does not require a reference image for its calculation, was determined to provide a quantitative assessment of the image sharpness. MSE, PSNR and SSIM were evaluated with the implementations available in scitik-image (https://scikit-image.org/), whereas scikit-video (http://scikit-video.org) was used for the calculation of VIPF. CPBD was determined by using the Python module available at https://ivulab.asu.edu/software/cpbd. For the super-resolution task, the quality of the generated images was determined by means of PSNR, SSIM, VIPF and CPBD. As a baseline for comparison with PSNR, SSIM and VIPF, the down-sampled images of 128×128 pixels were upscaled to 512×512 pixels by both linear and cubic resampling. The other image-to-image translations tasks were quantitatively assessed by means of MSE, PSNR and SSIM.

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 4 of 13
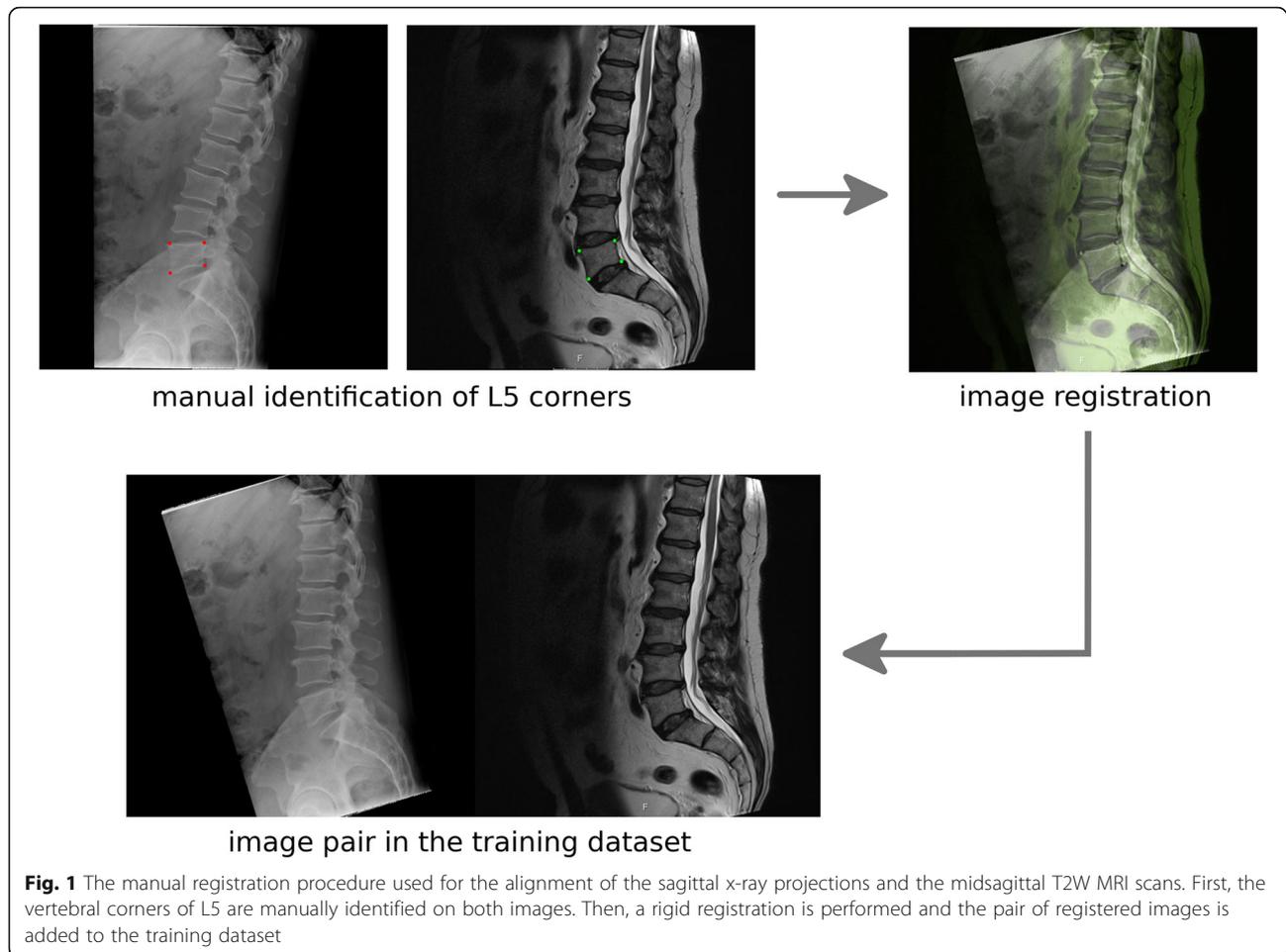
Fig. 1 The manual registration procedure used for the alignment of the sagittal x-ray projections and the midsagittal T2W MRI scans. First, the vertebral corners of L5 are manually identified on both images. Then, a rigid registration is performed and the pair of registered images is added to the training dataset

Clinical evaluation

A musculoskeletal radiologist with four years of experience evaluated the synthetic images of 30 anonymised and randomly selected patients. Specifically, the number of disc protrusions or herniations and the presence of artifacts were assessed in the images generated with super-resolution. The presence of L4-L5 degenerative disc disease, L5-S1 degenerative disc disease and L4-S1 Modic-type endplate changes [22] was assessed for the T1W-to-T2W translation and the reverse translation, as well as T2W-to-STIR and T2W-to-TIRM translation tasks. Finally, the presence of abnormal numbering and body compression fractures of the lumbar vertebrae were assessed on the x-ray to T2W translated images. After a delay of 14 days aimed at minimising the recall of given studies, the same musculoskeletal radiologist evaluated the target images (i.e. the native images) of the same patients. Specifically, the features previously evaluated in the synthetic images, except the super-resolution task-related artifacts, were assessed. Concordance between the synthetic and target images was calculated using κ-statistics and interpreted as follows: < 0.20, poor;

0.21−0.40, fair; 0.41−0.60, moderate; 0.61−0.80, good; 0.81−1.00, excellent [23].

Results

Conditional GANs improved image sharpness and visible details in the T2W images in the super-resolution task (Fig. 2). The generative model was able to add realistic details which were not visible in the low-resolution images. Additional visual information with a potential to be clinically relevant can be easily depicted in the intervertebral discs (including protrusions), in the spinal cord and even in the adipose tissue. However, the super-resolution images are affected, in a variable measure, by inaccuracies visually resembling truncation artifacts. Unexpectedly, the quantitative evaluation of the quality of the outputs, as compared with standard linear and cubic upsampling, did not result in a better performance of the GANs based on PSNR, SSIM and VIPF (Fig. 3). Nevertheless, the generative models were superior in restoring the image sharpness with respect to the ground truth, based on the assessment with the CPBD metric.

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 5 of 13
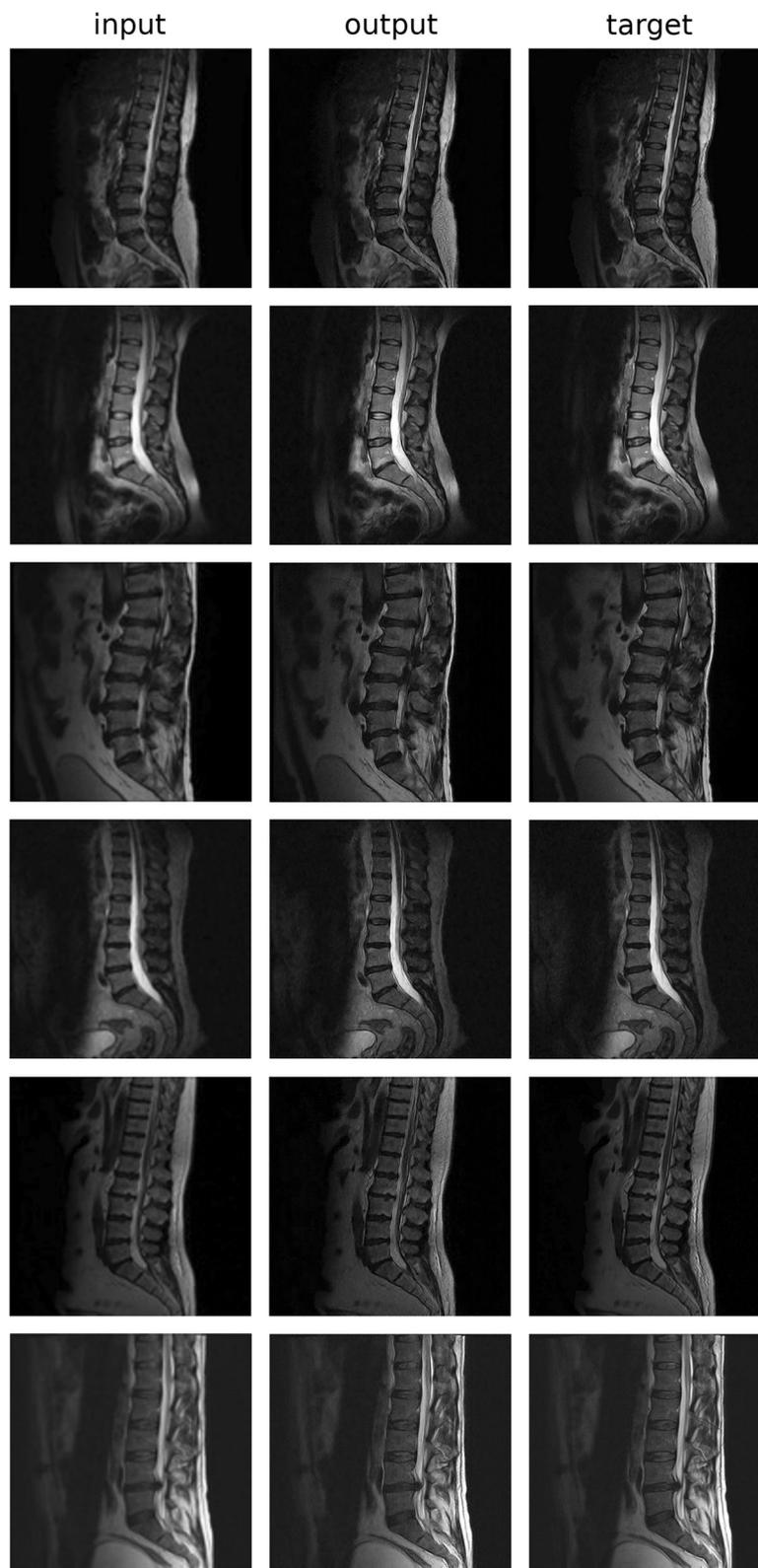
Fig. 2 Six examples of the use of GANs for the super-resolution task. input: original low-resolution image; output: output of the generative model; target: original high-resolution image. Multiple parallel edges similar to truncation artifacts are visible in proximity of high-contrast boundaries

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 6 of 13

Fig. 3 Quantitative comparison among the outputs of GANs (GANs), linear resampling (linear) and cubic resampling (cubic) by means of the metrics PSNR (**a**), SSIM (**b**), VIPF (**c**) and CPBD (**d**). For CPBD, the value of the metric calculated for the target image (target) is also shown for the sake of comparison
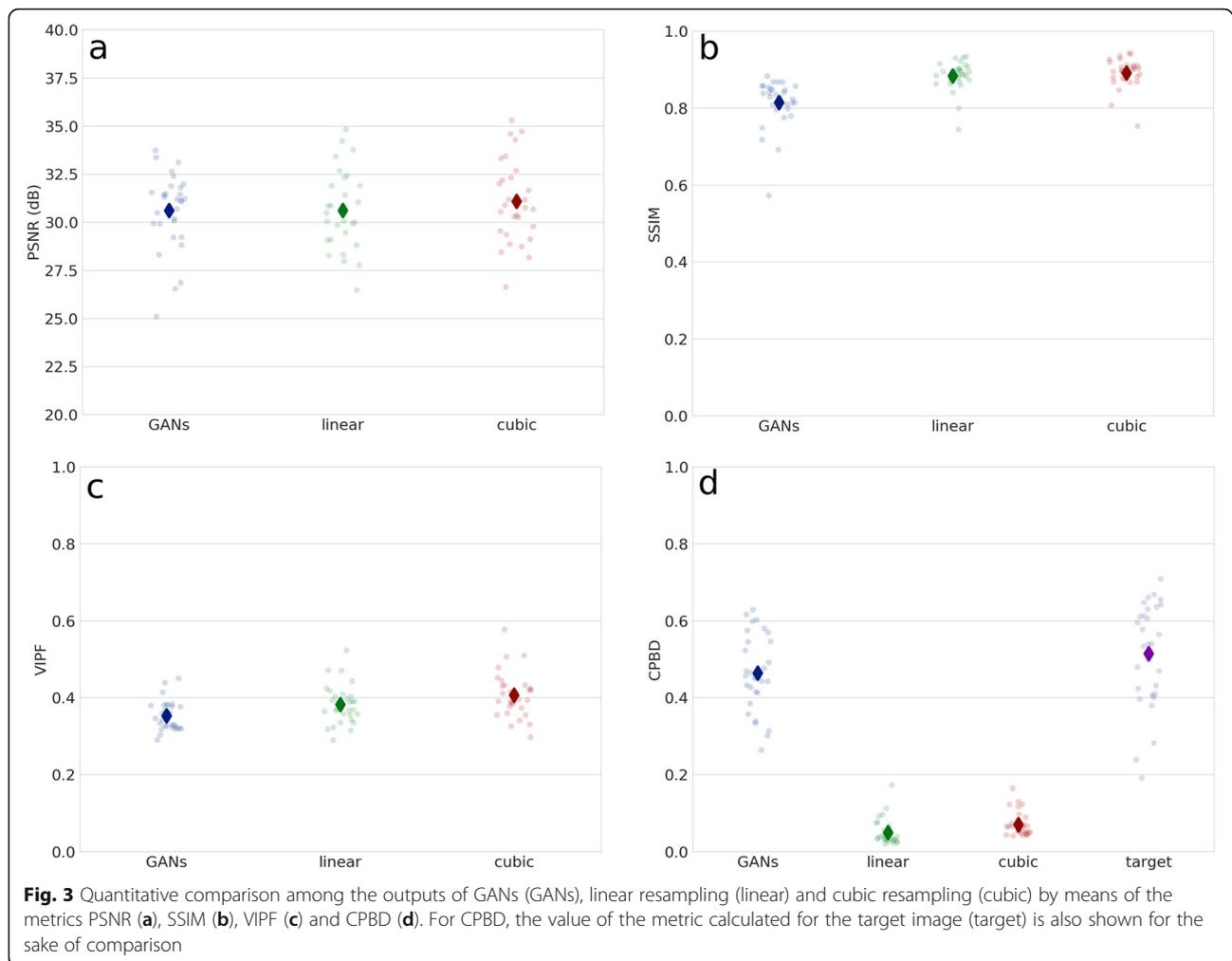
In the tasks related to virtual multimodal imaging, the performance of the conditional GANs can be judged as generally positive and promising; however, a deeper analysis of the outputs revealed several limitations. The translation between T1W and T2W images and vice versa (Fig. 4) demonstrated that the used generative model was able to deal well with the general features and differences between the two imaging modalities, such as the distinct grey levels in the spinal cord and in the nucleus pulposus of the intervertebral discs. However, healthy discs with high water content tended to have a slightly brighter representation in the nucleus, even in the synthetic T1W MR images, in contrast with the original images in which discs have a rather homogeneous grey level independently on the disc height or degeneration degree. On the other side, synthetic T2W images clearly depicted protruded discs and the correspondent compression of the spinal cord, which were less evident in the T1W images used as inputs. Degenerated discs tended to be correctly darker than healthy ones in synthetic T2W images. As expected, the

generative model was generally not able to deal correctly with Modic changes; in general, the signal alteration found in the original image was directly translated to the synthetic image, i.e. whereas type II changes were correctly represented, type I changes appeared similar to type III ones in the T1W-to-T2W translation.

From a qualitative point of view, the performance of the conditional GANs in translating T2W scans to TIRM and STIR scans appeared to be excellent (Fig. 5). Similar to the T1W-to-T2W translation, the model was capable of dealing well with the general features of the images, such as the fat suppression and signal alterations in the pathologic vertebrae.

The challenging translation from sagittal x-ray projections to T2W midsagittal MRI scans provided very interesting results which included a few evident errors in the depiction of the anatomy of the patient, but on the other hand highlighted the enormous potential of generative models (Fig. 6). Indeed, the synthetic MRI images showed in numerous cases incorrect anatomies, such as

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 7 of 13
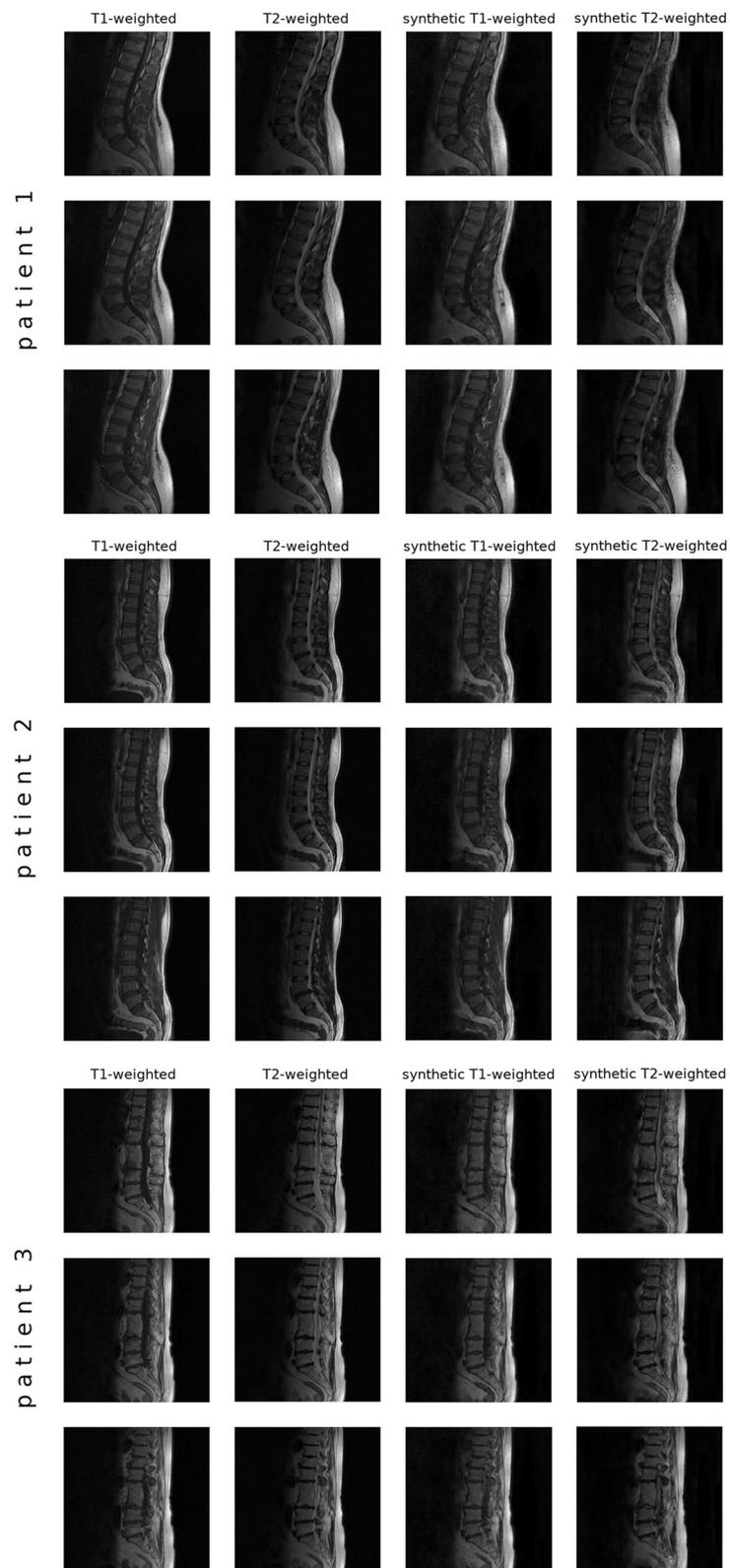
Fig. 4 Representative results (three slices for three different patients) of the translation from T1W to T2W MRI and vice versa. In the first patient, Type II Modic changes were correctly represented in both synthetic images, whereas for the Type I change in the third patient the synthetic T2W image showed a low signal instead of a high one. The L4-L5 disc protrusion of the second patient was accurately represented in the synthetic T2W image

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 8 of 13

Fig. 5 Representative results of the translation from T2W MRIs to TIRM (*left*) and to STIR (*right*) images. input: original T2W image; output: TIRM/STIR image created by the generative model; target: original TIRM/STIR image. Images from six exemplary patients are shown for each translation
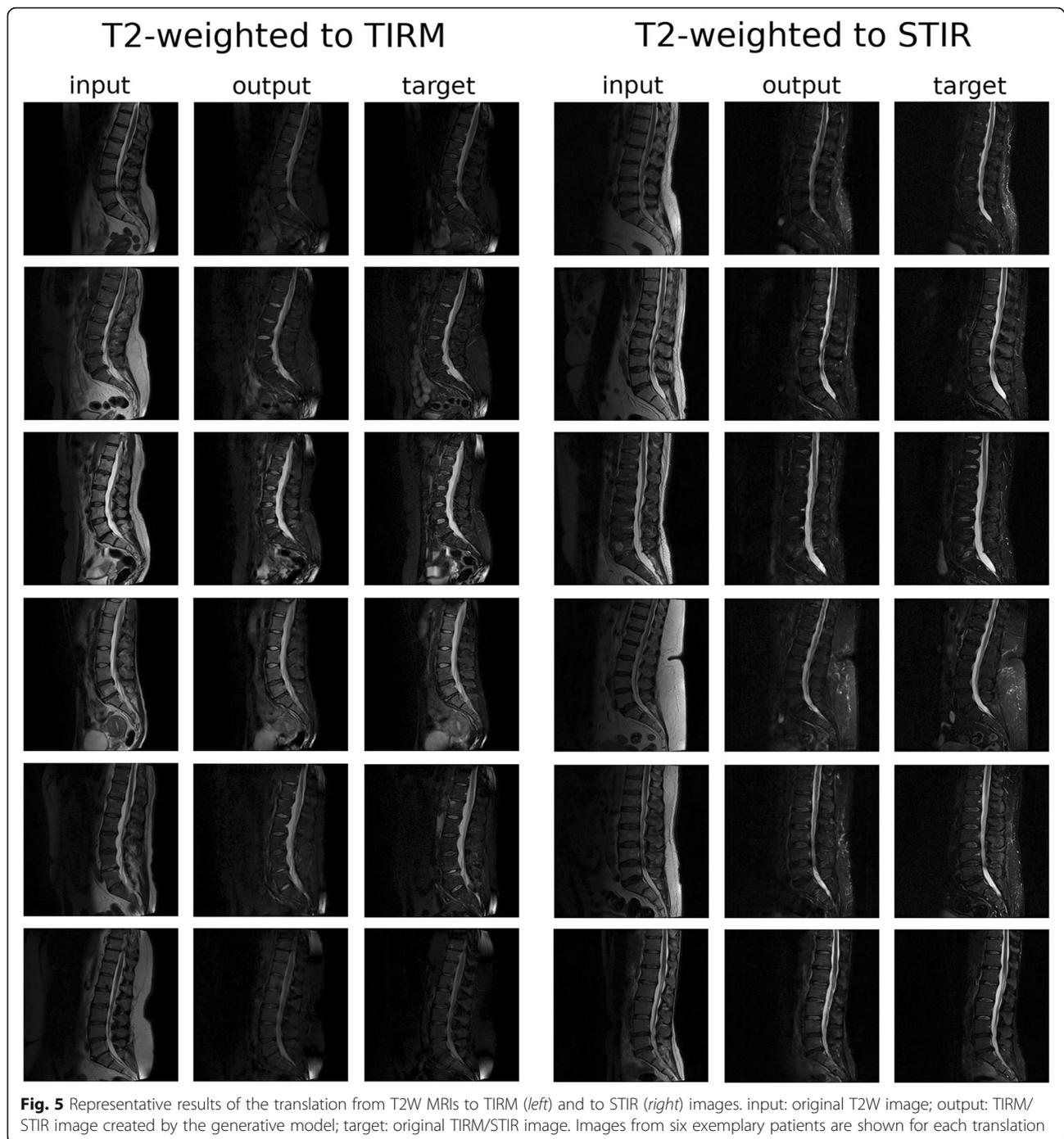
unrealistically long or short vertebral bodies, especially in regions not clearly visible on the x-ray projections, such as the thoracolumbar junction. Nevertheless, the conditional GANs were able to create realistically looking MR images which include a basically correct depiction of the spinal cord, of the intervertebral discs and of the layer of adipose tissue on the back. The lower lumbar lordosis due to the different posture (standing in the sagittal x-ray examination versus supine in MRI) was also realistically captured.

The quantitative evaluation of the image quality with respect to the ground truth revealed a consistent pattern among the three considered metrics MSE, PSNR and SSIM (Fig. 7). The translation of T1W images into T2W images and vice versa gave the most accurate results, followed by the translation from T2W to TIRM and STIR images respectively. As expected, the translation of sagittal x-ray projection to MRI consistently resulted in the worst quantitative performance.

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 9 of 13
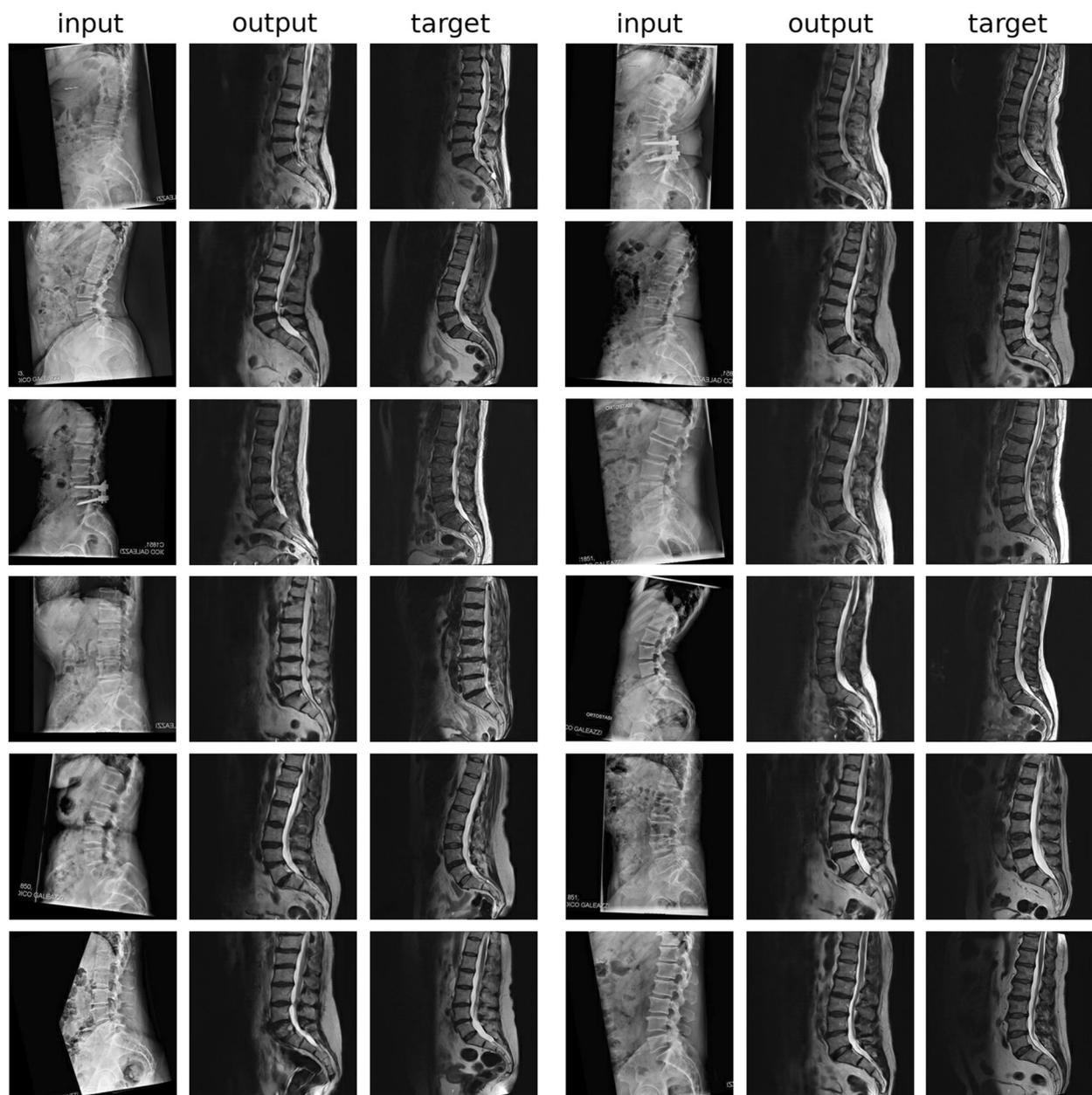
Fig. 6 Representative results (12 patients) of the translation from sagittal x-ray projections to T2W midsagittal MRI scans. input: original x-ray image, registered to match the alignment of the original MRI; output: synthetic T2W midsagittal MRI; target: original T2W midsagittal MRI. A few gross errors in the anatomy can be identified: the L1 in the first and fourth patients on the *left*; the sacrum in the first patient on the *right*; the thoracolumbar junction in the third patient on the *right*; L1 and the sacrum in the fourth patient on the *right*

With regard to clinical evaluation, the number of disc protrusions or herniations showed good concordance ($\kappa = 0.691$) between the native images and those generated with super-resolution. The median percentage of truncation artifacts amounted to 20%. Table 2 reports the concordance between the native images and those generated with virtual multimodal imaging. Moderate-to-excellent concordance was found when translating T2W to STIR and TIRM images, while the agreement was poor when translating x-ray projections into T2W images.

Discussion

In this paper, we explored the use of generative models, namely conditional GANs, for the creation of synthetic images of the spine and for the improvement of the quality of existing images. Despite several inaccuracies in the outputs, including some evident mistakes such as in the number of

Fig. 7 Quantitative evaluation of the quality, based on different metrics (MSE (**a**), PSNR (**b**) and SSIM (**c**)) of the various image-to-image translation tasks: super-resolution (SR), T1W to T2W MRI images (T1-T2), T2W to T1W images (T2-T1), T2W to STIR images (T2-STIR), T2W to TIRM images (T2-TIRM), sagittal standing x-ray projections to T2W images (Xrays-T2)
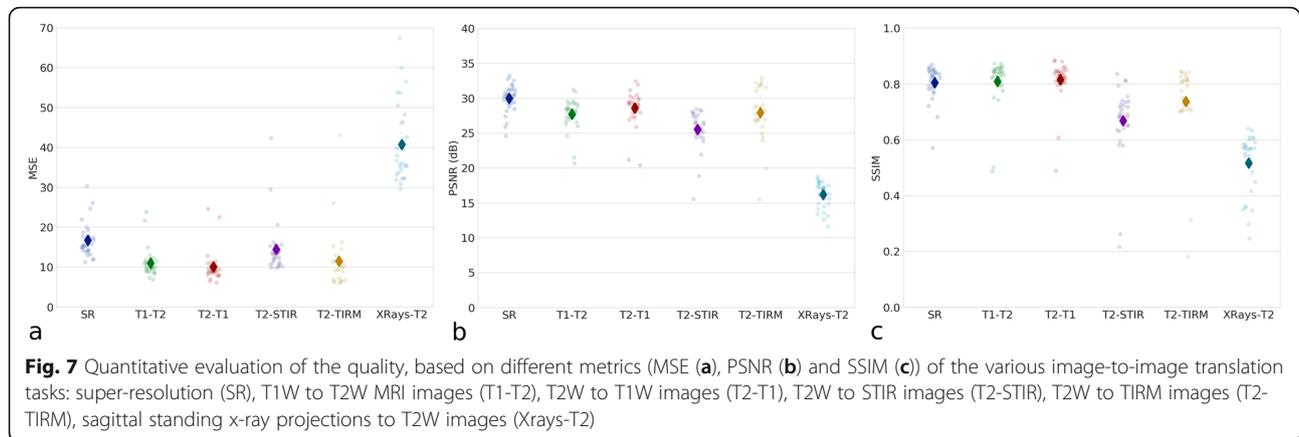
vertebrae, the general performance of the method should be judged as positive and very promising in light of future applications. As a matter of fact, no similar results have ever been reported in the available literature, neither with deep learning-based methods nor with other techniques. A paper describing a similar approach was aimed at the generation of synthetic images to be used in in-silico trials [13], rather than to proper clinical applications. Recent research highlighted the potential of conditional GANs for other radiological tasks, such as the improvement of the quality of low-dose positron emission tomography imaging [24] and noise reduction in low-dose computed tomography (CT) [25].

Taking into account the constant, fast advance of deep learning techniques for image synthesis and the large number of options for the technical refinement of the methods discussed in the present work, we foresee an enormous improvement of the quality of the generated data in the next future. It should be noted that the present work was not aimed at developing novel techniques for radiological image synthesis, but rather at exploring the potential of the methods currently available, knowing that research targeted specifically to radiology may provide even better results soon.

Although a preliminary quantitative assessment of the validity of the outputs of the generative models has been provided, the current work should still be intended as an exploratory proof of concept. As a matter of fact, the actual value of any innovative technique should be evaluated based on the impact that it can give on the practical applications rather than simply on the technical evaluation of its outputs such as the one here reported. Nevertheless, such a ground-breaking method opens new perspectives, in terms of potential applications, which still need to be explored.

Concerning spine imaging, possible clinically relevant uses include the grading of disc degeneration from planar x-ray imaging whenever an MRI scan is not available, the correction of the spinal shape due to different postures (e.g. standing versus supine), the improvement of the resolution of images acquired with low-field MRI scanners and the prediction of the effect of loading on the soft tissues, for example for the study of disc protrusions under loading. Additionally, virtual multimodal imaging may be ultimately integrated in PACS clients and Digital Imaging and COmmunications in Medicine (DICOM) viewers, to allow for a preliminary analysis of patients for which incomplete data are available. In case of diagnostic CT or MRI exams in which only a few slices have been acquired, the use of generative models may allow for synthetic re-slicing and thus high-quality visualisation also in the non-acquired orientations. Besides, we may foresee an MRI protocol for spine imaging using a single sequence or three sequences with different weighting composed by few slices. From that, such a model may generate a full set of MRI sequences, thus

Table 2 Concordance (expressed as κ-value) between the native images and those generated with virtual multimodal imaging

	L4-L5 disc disease	L5-S1 disc disease	L4-S1 Modic-type endplate changes
T1-weighted to T2-weighted	0.455 (moderate)	0.221 (fair)	0.406 (moderate)
T2-weighted to T1-weighted	0.086 (poor)	0.270 (fair)	0.286 (fair)
T2-weighted to STIR	0.842 (excellent)	1.000 (excellent)	0.592 (moderate)
T2-weighted to TIRM	0.842 (excellent)	0.933 (excellent)	0.691 (good)
	Lumbar vertebrae number	Lumbar vertebral body fractures	
Sagittal x-ray to T2-weighted	0.065 (poor)	0.051 (poor)	

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 11 of 13

remarkably reducing exam duration and MR system occupation. A similar approach has been recently reported on the knee, although based on a different technology [26].

For diagnostic purposes, the synthesis of a new image may not always be the optimal solution to achieve an improvement of the sensitivity and/or specificity of the diagnosis. Indeed, if the information required for the clinical evaluation is already available in the original image (e.g. information about intervertebral disc degeneration in a planar x-ray projection), the generation of a complete synthetic MRI scan showing the degenerative features of the disc may be deemed as superfluous for the diagnosis and grading of the disorder. Although we believe that several practical cases in which image synthesis can provide a clear benefit to musculoskeletal imaging, even from the clinical point of view, will definitely emerge as time goes by, it should be noted that simpler solutions may still be clinically advantageous for specific applications.

To our knowledge, the only application in which synthetic imaging data are nowadays used is MRI-only radiation therapy treatment planning [27]. In conventional radiation treatment, both MRI and CT images are acquired and used for planning and verification of the patient positioning. The simultaneous use of both imaging modalities requires a registration step, which introduces a systematic error not negligible from a clinical point of view. To eliminate it, an MRI-only workflow has been introduced, in which a synthetic CT is generated based on the MRI data. Various algorithms have been proposed for the generation of synthetic CTs, ranging from simple override techniques [28] to atlas-based ones [29, 30] and finally to sophisticated statistical models [31, 32]. The potential of conditional GANs for this specific application, possibly in combination with other consolidated approaches, is evident.

Although affected by artifacts, the super-resolution task provided very good results from a perceptual point of view. As a matter of fact, super-resolution is not a new concept and several algorithms have been proposed [33], with a special focus on MRI [34, 35]. Since the detection of small lesions may challenge even modern MRI scanners, this topic gains a specific clinical relevance. With respect to the classical MRI super-resolution techniques which rely on specific acquisition and reconstruction techniques, deep learning-based super-resolution can be applied as post-processing any time after the image reconstruction, with obvious advantages. Besides, generative models may add details not directly visible in the original images, based only on patterns found in similar patients, such as a specific shape, grey level or texture. The possible impact of these added details on the future clinical applications, either positive or negative, should not be neglected, since they can lead to misdiagnosis if they refer to non-existent pathological features. The clinical evaluation conducted in this study

highlighted that such artifacts indeed affected the outputs of the generative models, such as the number of vertebrae visible in the generated images and the occurrence of fractures in the translation from x-ray projection to T2W MRI. It should be noted that visual artifacts may be avoided or reduced by optimising the loss function of the model, for example by increasing the weight of the L1 regression with respect to the conditional GAN loss or by introducing a L2 loss term. Besides, such optimisation may benefit the quality metrics findings, whose results were not up to our expectations. As a matter of fact, the weights in the loss function used in this study arguably favoured sharpness over similarity to the target, with a clear negative impact on the metrics. These aspects were not investigated in the present paper, in which the weights of the two terms of the objective functions were kept fixed but need to be further analysed in future studies.

The results of image-to-image translation tasks also highlighted the potential of the generative framework. Similar to super-resolution, the novel methods can be applied in post-processing, since they do not require any modification to the acquisition and reconstruction stages. In this respect, generative models substantially differ from another documented MRI technique, synthetic MRI (SyMRI), providing a similar output, i.e. generating synthetic contrast-weighted images after the acquisition of the data [36–38]. Indeed, SyMRI dictates the use of a specific protocol creating a raw image which can then be post-processed to generate T1W, T2W and proton density maps and cannot be used on existing datasets acquired with other MRI protocols. It should be noted that, despite the generally convincing visual appearance of the translated images, a more extensive validation as well as an optimisation of the technique for the specific radiological applications are necessary before any clinical use of the novel techniques. The validation tests should address directly the specific clinical questions for which sequences such as STIR and TIRM are used, such as the diagnosis of soft-tissue tumours [39] and osteomyelitis [40], rather than being limited to a general evaluation of the quality of the synthetic images.

Due to its preliminary nature and its novelty, the present work suffers from several limitations, the most important of which is indeed the limited extent of the clinical validation. Furthermore, we decided to use an available implementation aimed to general image-to-image translation, without customising it to the specific application. As mentioned above, even simple optimisations such as the adjustment of the weights in the loss function may have a positive impact on the quality of the results. Another limitation pertains to the limited size of the training datasets, which has been constrained by practical issues related to the availability and traceability of the images. We expect that increasing the number of images constituting the training data would involve a major improvement in the quality of the outputs.

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 12 of 13

Conclusions

In conclusion, this proof of concept study showed that conditional GANs are able to generate perceptually convincing synthetic images of the spine, in super-resolution and image-to-image translation tasks. With respect to other methods providing analogous outputs, conditional GANs do not require specific acquisition and reconstruction techniques, and they can be employed in post-processing to any existing images. Although a clinical validation is still missing, we believe that conditional GANs, and deep learning-based generative methods in general, have the potential to be an upcoming innovation in musculoskeletal radiology.

Abbreviations
CPBD: Cumulative probability of blur detection; CT: Computed tomography; GANs: Generative adversarial networks; MRI: Magnetic resonance imaging; MSE: Mean square error; PSNR: Peak signal-to-noise ratio; SSIM: Structural similarity; STIR: Short tau inversion recovery; T1W: T1-weighted; T2W: T2-weighted; TIRM: Turbo inversion recovery magnitude; VIPF: Visual information fidelity in pixel domain

Authors' contributions
FG: drafting the paper, software development, creation of the training datasets, quantitative validation, data analysis; TB and GC: creation of the training datasets, data analysis, revision and approval of the paper; SG and EZ: clinical and quantitative validation, data analysis, revision and approval of the paper; FC and LMS: data analysis, revision and approval of the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Written informed consent for the use of the data for research purposes was obtained and double anonymisation of patients' data was performed. Ethical committee approval for this retrospective study has been obtained and patients' informed consent was waived.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests..

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Laboratory of Biological Structures Mechanics, IRCCS Istituto Ortopedico Galeazzi, via Galeazzi 4, 20161 Milan, Italy. [2]Unit of Diagnostic and Interventional Radiology, IRCCS Istituto Ortopedico Galeazzi, via Galeazzi 4, 20161 Milan, Italy. [3]Department of Neurosurgery, Humanitas Clinical and Research Hospital, Via Manzoni 56, 20089 Rozzano, Italy. [4]Department of Biomedical Sciences for Health, Università degli Studi di Milano, via Carlo Pascal 36, 20133 Milan, Italy.

References
1. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2:1
2. Deng L, Yu D (2014) Deep learning: methods and applications. Found Trends Signal Process 7:197–387
3. Salakhutdinov R (2015) Learning deep generative models. Annu Rev Stat Appl 2:361–385
4. Isola P, Zhu J, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. arXiv preprint. arXiv:1611.07004
5. Goodfellow I (2016) NIPS 2016 tutorial: generative adversarial networks. arXiv preprint. arXiv:1701.00160
6. Liu Y, Qin Z, Luo Z, Wang H (2018) Auto-painter: cartoon image generation from sketch by using conditional generative adversarial networks. Neurocomputing 311:78–87
7. Galbusera F, Bassani T, Costa F, Brayda-Bruno M, Zerbi A, Wilke HJ (2016) Artificial neural networks for the recognition of vertebral landmarks in the lumbar spine. Comput Methods Biomech Biomed Eng Imaging Vis 6:447–452
8. Jamaludin A, Lootus M, Kadir T et al (2017) ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. Eur Spine J 26:1374–1383
9. Koh J, Chaudhary V, Dhillon G (2012) Disc herniation diagnosis in MRI using a CAD framework and a two-level classifier. Int J Comput Assist Radiol Surg 7:861–869
10. Azimi P, Benzel EC, Shahzadi S, Azhari S, Mohammadi HR (2014) Use of artificial neural networks to predict surgical satisfaction in patients with lumbar spinal canal stenosis. J Neurosurg Spine 20:300–305
11. Scheer JK, Smith JS, Schwab F et al (2017) Development of a preoperative predictive model for major complications following adult spinal deformity surgery. J Neurosurg Spine 26:736–743
12. Oh T, Scheer JK, Smith JS et al (2017) Potential of predictive computer models for preoperative patient selection to enhance overall quality-adjusted life years gained at 2-year follow-up: a simulation in 234 patients with adult spinal deformity. Neurosurg Focus 43:E2
13. Galbusera F, Niemeyer F, Seyfried M et al (2018) Exploring the potential of generative adversarial networks for synthesizing radiological images of the spine to be used in in silico trials. Front Bioeng Biotechnol 6:53
14. Ledig C, Theis L, Huszár F et al (2016) Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint. arXiv:1609:04802
15. Li C, Wand M (2016) Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9907. Springer, Cham, pp 702–706
16. Zhu J, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593
17. Larsen ABL, Sønderby SK, Larochelle H, Winther O (2015) Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300
18. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham, pp 234–241
19. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13:600–612
20. Gu K, Zhai G, Yang X, Zhang W (2012) An improved full-reference image quality metric based on structure compensation. In: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, Hollywood, CA. IEEE, Piscataway, pp 1–6
21. Narvekar ND, Karam LJ (2011) A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). IEEE Trans Image Process 20:2678–2683

Galbusera *et al. European Radiology Experimental* (2018) 2:29

Page 13 of 13

22. Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR (1988) Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 166:193–199

23. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

24. Wang Y, Yu B, Wang L et al (2018) 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. Neuroimage 174:550–562

25. Wolterink JM, Leiner T, Viergever MA, Isgum I (2017) Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans Med Imaging 36: 2536–2545

26. Garwood ER, Knoll F, Hammernik K et al (2017) Accelerated knee MRI with machine learning based reconstruction compared to standard knee MRI protocol: analysis of protocol interchangeability and image quality. Paper presented at Radiological Society of North America annual meeting, Chicago, IL, November 28th 2017 (abstract SSG09-03)

27. Johnstone E, Wyatt JJ, Henry AM et al (2018) Systematic review of synthetic computed tomography generation methodologies for use in magnetic resonance imaging-only radiation therapy. Int J Radiat Oncol Biol Phys 100: 199–217

28. Schad LR, Blüml S, Hawighorst H, Wenz F, Lorenz WJ (1994) Radiosurgical treatment planning of brain metastases based on a fast, three-dimensional MR imaging technique. Magn Reson Imaging 12:811–819

29. Dowling JA, Sun J, Pichler P et al (2015) Automatic substitute computed tomography generation and contouring for magnetic resonance imaging (MRI)-alone external beam radiation therapy from standard MRI sequences. Int J Radiat Oncol Biol Phys 93:1144–1153

30. Uh J, Merchant TE, Li Y, Li X, Hua C (2014) MRI-based treatment planning with pseudo CT generated through atlas registration. Med Phys 41:051711

31. Johansson A, Karlsson M, Yu J, Asklund T, Nyholm T (2012) Voxel-wise uncertainty in CT substitute derived from MRI. Med Phys 39:3283–3290

32. Gudur MSR, Hara W, Le Q, Le QT, Wang L, Xing L, Li R (2014) A unifying probabilistic Bayesian approach to derive electron density from MRI for radiation therapy treatment planning. Phys Med Biol 59:6595

33. Van Reeth E, Tham IW, Tan CH, Poh CL (2012) Super-resolution in magnetic resonance imaging: a review. Concepts Magn Reson, 40A:306–325

34. Peled S, Yeshurun Y (2001) Superresolution in MRI: application to human white matter fiber tract visualization by diffusion tensor imaging. Magn Res Med 45:29–35

35. Tieng QM, Cowin GJ, Reutens DC, Galloway GJ, Vegh V (2011) MRI resolution enhancement: how useful are shifted images obtained by changing the demodulation frequency? Magn Res Med 65:664–672

36. Hagiwara A, Warntjes M, Hori M et al (2017) SyMRI of the brain: rapid quantification of relaxation rates and proton density, with synthetic MRI, automatic brain segmentation, and myelin measurement. Invest Radiol 52:647–657

37. Andica C, Hagiwara A, Hori M et al (2017) Automated brain tissue and myelin volumetry based on quantitative MR imaging with various in-plane resolutions. J Neuroradiol 45:164–16836

38. Drake-Pérez M, Delattre BMA, Boto J et al (2018) Normal values of magnetic relaxation parameters of spine components with the synthetic MRI sequence. AJNR Am J Neuroradiol 39:788–795

39. Tokuda O, Hayashi N, Matsunaga N (2004) MRI of bone tumors: fast STIR imaging as a substitute for T1-weighted contrast-enhanced fat-suppressed spin-echo imaging. J Magn Reson Imaging 19:475–481

40. Hauer MP, Uhl M, Allmann K et al (1998) Comparison of turbo inversion recovery magnitude (TIRM) with T2-weighted turbo spin-echo and T1-weighted spin-echo MR imaging in the early diagnosis of acute osteomyelitis in children. Pediatr Radiol 28:846–850