REVIEW ARTICLE
https://doi.org/10.1038/s41551-022-00914-1

nature biomedical engineering

Self-supervised learning in medicine and healthcare

Rayan Krishnan[1], Pranav Rajpurkar[2,4] ✉ and Eric J. Topol[3,4]

The development of medical applications of machine learning has required manual annotation of data, often by medical experts. Yet, the availability of large-scale unannotated data provides opportunities for the development of better machine-learning models. In this Review, we highlight self-supervised methods and models for use in medicine and healthcare, and discuss the advantages and limitations of their application to tasks involving electronic health records and datasets of medical images, bioelectrical signals, and sequences and structures of genes and proteins. We also discuss promising applications of self-supervised learning for the development of models leveraging multimodal datasets, and the challenges in collecting unbiased data for their training. Self-supervised learning may accelerate the development of medical artificial intelligence.

Medical artificial intelligence (AI) has been driven by advancements in deep learning and in the creation of datasets. Algorithms for medical AI have been developed on medical tasks intended to diagnose, predict and recommend treatments across a variety of medical modalities and data types, such as electronic health records (EHRs), chest X-rays, electrocardiograms and protein sequences[1]. When building algorithms for medical AI, a central challenge is their reliance on the availability of annotated input data at scale, often in the hundreds of thousands, if not millions, of datapoints. Addressing this bottleneck would enable the development of accurate AI algorithms for a much broader range of tasks in health and disease, from diagnostics to monitoring to treatment decisions. In this Review, we highlight recently developed and promising sets of techniques in self-supervised learning, and their challenges and opportunities when used in medicine and healthcare.

Deep learning is the dominant approach for developing medical AI. However, the success of its applications relies heavily on the availability of annotated datasets. Deep-learning models are typically trained using a supervised-learning paradigm, where the models learn to map an input (such as a chest X-ray image or a health record) to an output (for example, a diagnosis of pleural effusion, or the prediction of myocardial infarction). For the models to learn relevant patterns in the data, training them via supervised learning requires large datasets in which each input is annotated with its corresponding output. However, much more emphasis has been placed on building and testing models than on the heavy-lifting work of building annotated datasets. This is partly because, for most medical tasks, building the required large datasets would prove inordinately expensive[2]. Still, there has been insufficient commitment to expand the resources needed to create such annotated datasets. For common image types, such as chest X-ray images, images of skin lesions, retinal photographs and brain computed-tomography scans, the existing datasets have been repeatedly used.

Medical datasets carefully annotated by experts are hard to create at scale. Non-medical deep-learning models have been incredibly successful when trained on ImageNet, which harnessed 49,000 Mechanical Turk workers (Amazon's Mechanical Turk is a crowdsourcing marketplace for outsourcing tasks that typically require human intelligence) and hundreds of academics and citizen scientists to label approximately 15 million images of 21,000 classes (such as 'broccoli' and 'hummingbird')[3]. However, the labelling of medical datasets requires experts and considerable time. Interpreting a medical image, such as a tissue slide or electrocardiography (ECG) data, typically demands even more time per image than the labelling of natural objects or other diagnostic data used in clinical practice. For instance, chest X-ray images in the CheXpert dataset[4] were labelled at an estimated rate of 2–5 min per study, whereas image samples from ImageNet were labelled at an average rate of 50 images per minute[3]. In addition, methods for automated labelling—which have enabled weakly supervised learning, a technique that leverages noisy or imprecise sources to alleviate the burden of obtaining hand-labelled datasets—often require domain expertise and substantial development time. A related challenge is the vital need of datasets to be comprehensive and to fully represent the diversity of the data (in particular, of the relevant pathologies and patients).

Because medical domains are difficult to label, one method to build more capable models is to train them on a large and general dataset such as ImageNet, and then re-train the models on the smaller and specific medical task. In many applications, such a process of transfer learning (from a general domain to a specific domain) has allowed models to perform better than those trained from scratch[5]. However, applying transfer learning to the training of models for medical AI has an inherent problem: the first training task is typically not medically relevant, so the characteristics that the model learns may not be valuable for the medical task. Self-supervised learning is a better method for the first phase of training, as the model then learns about the specific medical domain, even in the absence of explicit labels.

Unlike labelled datasets, which are difficult to create, unlabelled medical data are plentiful. These include images, gene or protein sequences, electronic health records and pathology slides that have been collected from patients but not explicitly labelled with a diagnosis. Unlabelled datasets can be leveraged to build self-supervised models that learn complex structures in the data via new

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Biomedical Informatics, Harvard University, Boston, MA, USA. [3]Scripps Research Translational Institute, Scripps Research, La Jolla, CA, USA. [4]These authors contributed equally: Pranav Rajpurkar, Eric J. Topol.
✉e-mail: pranav_rajpurkar@hms.harvard.edu

supervised-learning tasks; for example, by occluding portions of the data and expecting the model to predict what has been hidden, or by providing two samples from the same patient and training the model to associate them strongly. After such preliminary training, these general models can be trained again on a much smaller set of labelled examples for the final medical task.

In such a first phase of training, the models are trained on a preparation task, referred to as a 'pretext task'. Because the data used for such pretext tasks are unlabelled, the trained model cannot yet solve the primary task. Instead, after being trained on a pretext task, the model, referred to as a 'featurizer', can take an input sample and output a vector of numbers that represent the important aspects of the input in a machine-readable form. This pretext training phase is valuable because the model can learn how to find useful features or attributes in the data, even before seeing any labelled data. In the second phase of training, the featurizer is trained on a dataset of explicit labels. This enables the model to incorporate its knowledge of the data to perform the relevant medical task. In what follows, we highlight specific examples of self-supervised pre-training tasks across medical domains and data types in its two major forms: contrastive learning and generative learning[6,7].

Contrastive learning

The primary objective of pre-training a model with contrastive learning is to make the model associate similar samples and dissociate dissimilar samples. The task for the model consists of predicting whether a pair of samples are positive pairs (hence, are closely related) or negative pairs (and thus unrelated). The data may already be naturally structured for this setup.

For example, a face-detection dataset may have many photos of each subject's countenance. In this case, photographs of the same person would make positive pairs, and photographs of different people would constitute negative pairs. Ideally, the featurizer should take two images as input, and produce two corresponding vectors that are numerically similar to one another if the images are positive pairs (and numerically far from one another if the images are negative pairs)[6–8]. By receiving many pairs and predicting whether they are positive or negative, the model becomes a strong featurizer, increasingly recognizing relevant information in the images. When the model has achieved a sufficiently comprehensive 'understanding' of the properties of the data, it is then trained explicitly on the available set of labelled data, where the input is a particular face and the expected output is a name.

Case studies. The most widely used strategy to generate training data for the contrastive-learning pretext task involves data-augmentation techniques. Data augmentation is the process by which minor changes to the data can be made to create new samples[6,9]. For example, images can be slightly cropped or rotated to create a distinct image that has the same content as the original image, or reflected to create a new corresponding image (Fig. 1). Because the subject and content of the image remain the same, altered pairs of the same original image are positive pairs, whereas altered versions of different original images are negative pairs. Depending on the data used, different augmentation techniques can be applied. For example, contrastive-learning augmentation techniques can also be applied to diagnose heart and lung diseases from digital stethoscope data[10]. This methodology has also shown its potential in work that has applied a version of contrastive learning to chest X-ray diagnosis[9,11–13]. In what follows, we highlight how this technique can be combined with other data variants. First, we describe applications for medical-image and sensing modalities, and then for sequence data or structured data, such as molecules and DNA sequences.

When multiple scans or samples are taken from the same patient, a multiple-viewpoint method can be applied. These multiple views can be used directly as positive pairs for the contrastive-learning task. One study[14] worked with chest X-ray data with both lateral and frontal scans and with dermatology data with photos from different perspectives (where only a single view was available, a data-augmentation step was used to generate image pairs; where two views existed, a minor data-augmentation approach was used on the distinct frames). Another study showed that data augmentation using multiple views of the same patient performed best for the generation of positive pairs for training diagnostic models based on chest X-ray images[15].

Self-supervised learning is particularly applicable for classification tasks based on histology slides, which are particularly challenging to annotate. In histology-based diagnosis, every sample contains a microscopy image of many cells, each of which may be cancerous. Each sample contains many cells, and the sample is considered cancerous if at least one cell in the image exhibits cancerous features. Traditionally, data have been labelled for this task by having an expert painstakingly annotate regions of pixels according to whether the regions contain cancerous or non-cancerous cells. Instead, unlabelled data can be used to create a self-supervised pre-training task by selecting sub-patches from the slide to use as positive pairs[16–18]. This relies on the patch of cells following the property that if one cell exhibits a label, all instances in the window share that same label. This methodology was developed further by using a self-supervised knowledge-distillation process to build an effective 'student' model that could classify lower-resolution images[19].

In addition to its applicability to spatially structured data, self-supervised learning can also be applied to data collected across time. For example, ECG data (if collected from multiple leads) are both spatially and temporally structured, and in conjunction with data-augmentation techniques have allowed models to generate spatiotemporal associations for the same underlying patient and disease[20,21]. In fact, physiologically rooted data-augmentation techniques can offer superior pre-training performance[22]. Because videos have both spatial and temporal variations (the contents of a single frame belong to the same space, and frames are collected across time), self-supervised learning has also been applied to natural videos and more recently to medical videos. For instance, a model for the interpretation of ultrasound videos was pre-trained in two phases[23]. In the first phase, video frames were shuffled, and the model was trained to predict the correct ordering of the frames to support the model's 'understanding' of temporality. In the second phase, a frame underwent a geometric transformation (such as rotation or translation), and the model was trained to predict the original frame.

These techniques of data augmentation (also when using multiple views) can be applied to other data types and tasks, such as the discovery of small-molecule drugs[24]. The task of using molecular structure to predict molecular properties is constrained by the limited number and size of labelled datasets and by the large space of molecules and molecular structures. By assuming that molecules with a similar structure have similar properties, one can create a positive pair of structures by applying a data-augmentation technique that randomly removes some atoms in the structure to create a modified version of the original. Molecules can also be augmented by randomly removing some bonds, or by removing whole portions of atoms and bonds from parts of the molecule[25]. For protein structures, similar strategies have been used; specifically, predicting the spatial distance between acids in the protein as well as missing edges in the structure[26]. Notably, a graph neural network trained on 10 million unlabelled molecules was more interpretable than previous models, and generated clear representations of molecules by topographical structure and functional groups[26].

Self-supervised techniques have also been applied to DNA-sequencing data[27] to determine whether DNA fragments align, as well as differences in the number of mutations irrespective of alignment. To do this, self-supervised learning can be applied by training

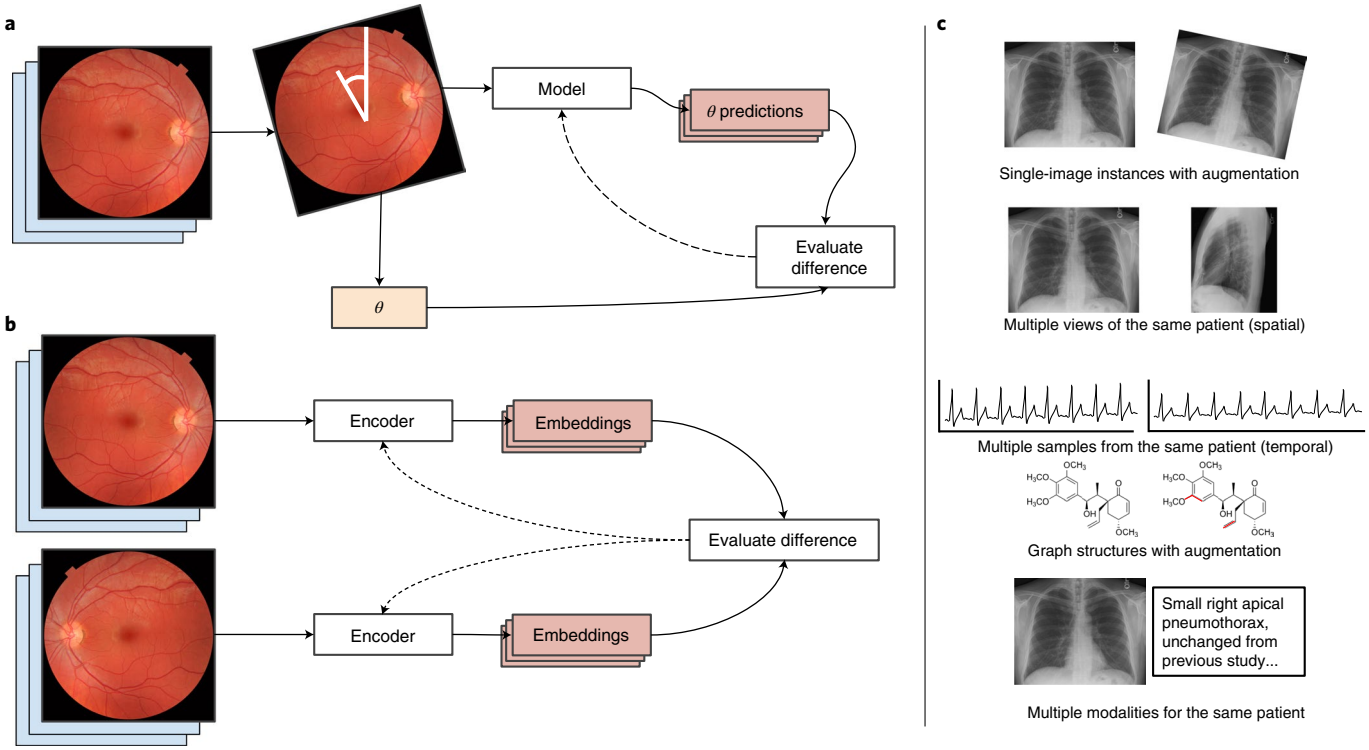

Fig. 1 | Contrastive learning. a, Example of a self-supervised pretext task, in which a fundus retinal image is rotated and the self-supervised learning model learns to predict the original image by evaluating the difference in rotation angle θ. **b**, A pretext task in which a fundus retinal image is augmented via horizontal reflection. An encoder model is trained to produce similar embeddings (representations of discrete variables as continuous vectors) across the two images. **c**, Examples of positive data pairs used in contrastive learning. From top to bottom: two rotations of the same X-ray image; two spatial views of the same patient; two time points of an ECG trace; two chemical structures of a small molecule differing slightly in chemical bonding (red); two data modalities: X-ray image and a textual description of it.

a model to predict a known number of mutations applied to a reference sequence; the model can then be used to predict the pairwise identity scores of sequences.

Generative learning

In generative pre-training, defining labels allows for the application of supervised-learning techniques to self-supervised learning. For instance, Wikipedia does not contain any definitions for the words in the corpus, and hence does not have any labels for individual words. However, because sentences are structured as an ordered sequence of words, such structure can be used to frame a new supervised-learning task where words in a sentence are randomly blanked out (masked) and the model is expected to use the context of the surrounding words to predict the masked word (Fig. 2). The masked word can then be re-obtained from the original text (and, therefore, there is an explicit 'right answer' that the model can be trained to predict). In fact, this procedure of learning the meaning of words via contextual clues is one way that we all use to guess the meaning of new words. A model that can successfully predict the missing word must also 'understand' its meaning (and hence, it can effectively 'read'), and this ability can be used to perform a specific task, such as identifying key topics in a piece of text.

Case studies. Text-masking techniques can be easily extended to tasks in the medical domain, such as tasks involving EHRs or protein sequences (which can also be treated as text)[26,28–32]. Time-series data (such as electroencephalography scans) can be used to frame a pre-training problem for the prediction of the next period (such as the next waveform)[33,34]. In addition, images, which can be pre-trained via contrastive loss, can also be used for generative tasks[35] by blocking out parts of the image and training the model to fill what was removed.

Masked-word pre-training (guessing the hidden word via context words) has been extensively used after the development of transformers—a state-of-the-art class of neural-network architectures for language tasks[36]. These models work by taking a sequence of tokens (individual words in sentences, for example), identifying the connections between each token, constructing a representation of the sequence and using the representation to produce an output sequence. Transformers have recently been applied to non-textual data, including images and protein structures. Notably, a transformer-based design was used to build models that can take arbitrary input data to learn representations of the data[37–39]. A model might do this by using one modality of input to filter or augment the information available in another data modality (this is known as 'cross attention'). These multimodal featurizers can be used in downstream tasks that make use of more than one data modality as input.

A direct application of generative self-supervised learning is the parsing of EHRs. Masked-word pretext tasks can be applied to EHRs because they contain textual information (background information of patients, such as gender, weight and height, as well as information recorded by doctors during patient visits). Models trained to make sense of EHRs can be used to predict the likelihood of patients developing diseases, which may enable more accurate early-disease detection[28,30,31]. For instance, a model adapted from the model known as BERT (for bidirectional encoder representations from transformers) to work with EHRs (the BEHRT model) was trained on a dataset of 1.6 million patients to predict disease likelihood for 301 pathologies[28] (the more recent Med-BERT model

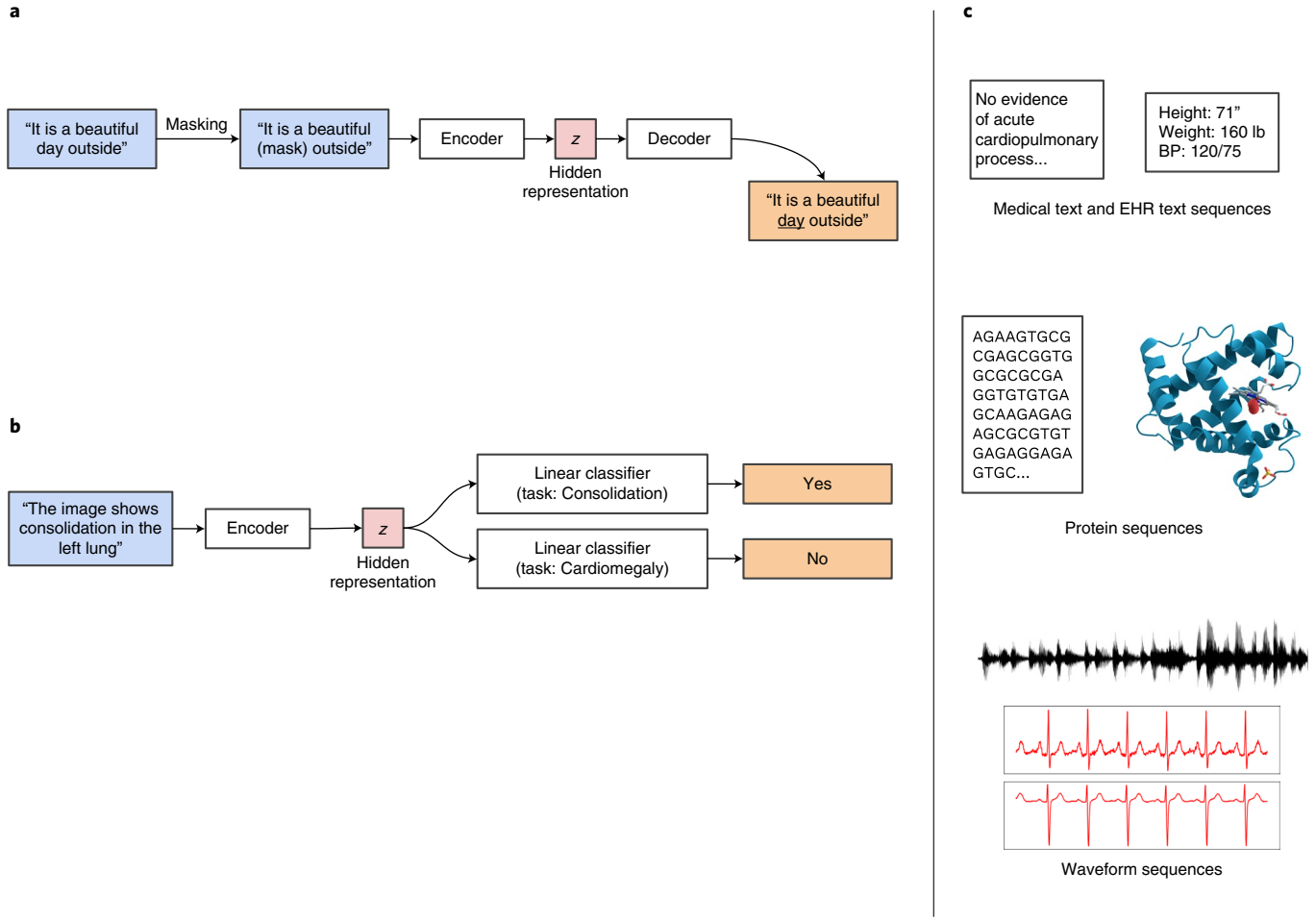

Fig. 2 | Generative pre-training. a, Example of pre-training with a masked-language model. The original text sequence has a random word masked, and the model is trained to restore the masked word. An encoder–decoder model is used to generate a 'hidden' numerical representation (z) of the data. This can also be performed on medically relevant text and on other data types. **b**, The trained word-guessing model depicted in **a** can be used to solve particular medical tasks, such as disease detection from text. The blue, red and orange backgrounds denote inputs, vector representations (embeddings) and model predictions, respectively. **c**, Examples of data types for which generative self-supervised learning has been applied. From top to bottom: free text and structured text from EHRs; protein sequences and structures; bioelectric waveforms from the brain (from electroencephalography) and the heart (ECG traces).

expanded the training dataset to 28 million patients[30]). These models could be trained on such large datasets because they were structured as text sequences for self-supervised masked pre-training (the data did not need to be manually labelled).

Generative language models can also be applied to protein sequences. Sequences are easy to collect, but it is difficult to label them for structure and function. Techniques for natural language processing have been applied directly as a pre-training task for protein data by framing them as supervised tasks for predicting the structure or functions of proteins with partially masked sequences[40,41]. During training, such models randomly mask parts of a hidden representation embedding and predict what was missing[42]. In addition to using generative techniques, a self-supervised contrastive-learning task can be constructed by creating positive protein fragments taken from the same protein sequence, as well as negative pairs from fragments from different sequences[43]. Such a pre-trained model can then be applied to other tasks, such as predicting a protein's secondary structure, stability or function.

Benefits and limitations

Self-supervised learning is primarily limited by the difficulty in finding and selecting useful pre-training tasks. There is no indication that applying a language-masking method will necessarily result in a high-performing model. For contrastive learning, the performance of the model is highly dependent on the data-augmentation technique used. Existing work is mostly empirical, relying on testing many different pretext tasks to identify which tasks are most useful. In some cases, it seems that having domain knowledge about the data and their structure may suggest augmentation techniques that are relevant for the particular data type[15,44].

After self-supervised training, some models have required fewer labelled examples to reach the same performance than models trained only through supervised learning (for instance, a self-supervised model for chest X-ray images used less than 10% of the labelled data to match the performance of a supervised model[11]). The implication is that new medical diagnosis tasks with limited labelled samples could be addressed using machine learning. This is particularly meaningful for rare pathologies, for which building datasets is naturally more difficult.

Although applications of self-supervised learning to medical tasks are currently few, the models have in many cases met or exceeded the performance of their strongly supervised counterparts[45,46]. For example, using X-ray images and radiologist reports for pre-training, a self-supervised diagnosis model using only 1%

of the labelled data met or exceeded the performance of baseline models. Similarly, a self-supervised model for the detection of diabetic retinopathy using only 25% of the labelled data performed similarly to other supervised models[47]. Self-supervised pre-trained models require fewer labelled examples for training due to faster convergence[48]. Improvements in accuracy with limited labelled data have been attributed to the ability of self-supervised models to first learn what features to consider as meaningful differences. In fact, self-supervised models can learn to segment images into particular objects without any labels (as exemplified by the Distillation with NO Labels model[49] which, after training with paired image data without any segmentation labels, correctly identified pixels belonging to different object classes).

Multiple-instance learning and multiple-viewpoint learning have produced models that are not misled by minor changes in the data[14]. In these training paradigms, images with different perspectives of the same object or pathology are considered similar (positive pairs), and thus the erroneous differences in perspective are not considered to be important features. For instance, if models consider two photos of different perspectives as positive pairs, they may associate important features with the subject of the photos rather than with their perspective. Strongly supervised models are liable to shortcut learning[50] by associating erroneous distortions of facets of the image with the label. This is commonly seen in models that rely on image backgrounds more heavily to predict the subject[51]. Instead, self-supervised models trained using data augmentation can produce models that are less vulnerable to minor visual distortions. Hence, by using data-augmentation techniques (for images, rotation, cropping or brightness alterations, for example), the models can learn to avoid associations with these types of image variation. Although some supervised-learning methods may be robust to such distortions, they may not generalize that well to out-of-distribution tests. This lack of generalization has been effectively counteracted by more extensive augmentation, reinforcing the importance of data augmentation for the success of contrastive learning[52] (in early work using mammograms for the detection of cancerous lesions, self-supervised methods have generalized better[53]). However, generalization depends on the task of interest and data-augmentation strategy.

Forthcoming developments

The uses of self-supervised learning for tasks in medical diagnosis thus far are suggestive of forthcoming progress. One main advantage of self-supervised techniques is that they can more easily make use of multimodal data. For example, the AI research and deployment company Open AI has combined text and image models into one model (named CLIP, for contrastive language-image pre-training)[49] that can perform contrastive learning with the outputs of an image encoder and a text encoder. Another model (ConVIRT, for contrastive visual representation learning from text)[11] can learn diagnostic labels for pairs of chest X-ray images and radiology reports. Contrastive learning can also be used to learn correlations between chest X-ray images and their reports, to generate text reports for new X-ray images (the CXR-RePaiR model)[54]. Another model leveraged the labelling of retinal fundus images with retinal-thickness values (obtained via optical coherence tomography) to build positive pairs for the pre-training of a model for the diagnosis of diabetic retinopathy[52]. After pre-training, the model relied only on the retinal images to classify them according to the severity of diabetic retinopathy. Pre-training with retinal-thickness data allowed the model to better 'understand' the most salient aspects in the retinal images; this improved its performance on the diagnostic task over that of a model trained solely with the retinal images. We expect that data from multiple medical tests or from patient information in EHRs will be increasingly used to create high-performance models that leverage self-supervised learning.

Models trained with multimodal data may be created with two different goals. First, they may make use of multimodal data when they are deployed, which would make them more relevant in clinical settings. For example, if a patient's age, weight and blood pressure are recorded before performing an ECG exam, then the models can be trained to also make use of these data rather than be trained only on the ECG data. These two types of data are likely to contain different yet diagnostically useful information about the patient. Second, the models may be pre-trained with more extensive tests (not commonly performed at the time of diagnosis) to build featurizers that have a superior 'understanding' of the underlying disease. For example, the model that leveraged self-supervised learning for the diagnosis and classification of the severity of diabetic retinopathy used retinal-thickness measurements, which are not routinely collected in screening programs for the disease and thus were not available for the testing of the model. By using more expensive medical tests for pre-training, the self-supervised model learned relevant associations in the data (a more nuanced 'understanding' of how to interpret a fundus image; that is, a better representation of it) that may lead to performance improvements when using only the more limited tests available during model deployment.

Featurizers can also use labelled data to perform other tasks in the medical domain. For instance, models can be trained to cluster patients into subgroups or to use specific types of patient. For example, by using contrastive learning to build a featurizer for chest X-ray images, a featurizer model was trained to identify patient groups that are most at risk of deterioration, even when labelled data are limited[55]. Another self-supervised method was trained on pathology slides to predict whether patients will show a response to a particular treatment[56]. The development of machine-learning models to predict the need for intervention and patient responses to treatments is relatively underexplored (with respect to the development of models for diagnostic tasks). Patient-specific models may aid the development of applications in personalized healthcare.

Self-supervised learning may also allow for models that can interpret multimodal data across time. This includes multiple 'views' of a patient from different tests, as well as previous medical history and patient testimony. With the increased use of fitness trackers and the increasingly regular collection of health data, machine-learning models could help interpret extensive datasets that would be difficult for physicians to do on their own. Building models that rely on multimodal data will require compiling datasets that incorporate data from medical tests from many patients. Much of existing research in machine learning for healthcare revolves around improving diagnostic tasks in controlled settings (typically using retrospective datasets and comparing performance to those of individual experts), which is insufficient to demonstrate clinically relevant success for a model. Models relying on comprehensive multimodal data and self-supervised learning will probably lead to more reliable clinical implementations in real-world settings.

It is essential to consider the consequences of overreliance on previously collected datasets. Models trained on biased data are vulnerable to become biased themselves[57]. This has been seen in the creation of large language models trained on open-text corpuses. For instance, models trained to translate from a language that lacks gender-specific pronouns into English showed bias for genders by occupation (for example, 'he is an engineer' and 'she is a nurse'). In the medical domain, large language models have exhibited bias against females and minorities by recommending that pain medication not be prescribed[58]. In addition, gender-biased datasets yield models that perform better on the majority class[59]. To avoid perpetuating biases found in historical data, new data will need to be collected and scrutinized to meet high standards for quality.

The need for the continuous collection of new data suggests that new systems will need to be built to support data acquisition for self-supervised models. Labelling services (such as Amazon Mechanical Turk) have become popular for procuring labelled large datasets, but to produce high-quality unlabelled datasets, new procedures and services will need to be developed. To create models ready for deployment, a priority is to ensure that the models are trained on unbiased data that match the data distribution of the deployment setting. To support this, new data-collection procedures that are not biased against certain patient subgroups can be implemented at the hospitals where the model will be deployed. In some cases, it may be more effective to create independent models by patient subgroup. Moreover, workflows that facilitate the labelling of data by medical experts as the patient is studied and tested, rather than after the data are collected, could be developed. Furthermore, labels could also be given to self-supervised groupings of the data. For instance, after clustering on self-supervised representations, medical experts could characterize key visual features shared between examples belonging to the same cluster and that are different from examples in other clusters[60]. These ideas would enable scalable methods for the building of large and heterogenous sets of high-quality unbiased data.

Overall, by using large unlabelled datasets to pre-train machine-learning models, self-supervised learning improves the performance of downstream tasks. This is particularly relevant for training models to perform medical-diagnosis tasks for which large labelled datasets are difficult to procure. For contrastive-learning and generative-learning pretext tasks, data-specific augmentation strategies and techniques are needed. We postulate that, in many unexplored areas of medicine, self-supervised learning leveraging multimodal data will enable the creation of high-performing models that better 'understand' the underlying physiology.

Received: 23 November 2021; Accepted: 27 June 2022;
Published online: 11 August 2022

References

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
2. Sambasivan, N. et al. "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2021); https://doi.org/10.1145/3411764.3445518
3. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
4. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* 590–597 (AAAI Press, 2019).
5. Huh, M., Agrawal, P. & Efros, A. What makes ImageNet good for transfer learning? Preprint at https://doi.org/10.48550/arXiv.1608.08614 (2016).
6. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. 37th International Conference on Machine Learning* (eds. Daumé, H. III & Singh, A.) 1597–1607 (PMLR, 2020).
7. Chen, X., Fan, H., Girshick, R. & He, K. Improved baselines with momentum contrastive learning. Preprint at https://doi.org/10.48550/arXiv.2003.04297 (2020).
8. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow Twins: self-supervised learning via redundancy reduction. In *Proc. 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) 12310–12320 (PMLR, 2021).
9. Sowrirajan, H., Yang, J., Ng, A. Y. & Rajpurkar, P. MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models. In *Medical Imaging with Deep Learning 2021* 727–743 (PMLR, 2021).
10. Soni, P. N., Shi, S., Sriram, P. R., Ng, A. Y. & Rajpurkar, P. Contrastive learning of heart and lung sounds for label-efficient diagnosis. *Patterns* **3**, 100400 (2022).
11. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. Preprint at https://doi.org/10.48550/arXiv.2010.00747 (2020).
12. Sriram, A. et al. COVID-19 prognosis via self-supervised representation learning and multi-image prediction. Preprint at https://doi.org/10.48550/arXiv.2101.04909 (2021).
13. Han, Y., Chen, C., Tewfik, A. H., Ding, Y. & Peng, Y. Pneumonia detection on chest X-ray using radiomic features and contrastive learning. In *2021 IEEE 18th International Symposium on Biomedical Imaging ISBI* 247–251 (IEEE Computer Society, 2021).
14. Azizi, S. et al. Big self-supervised models advance medical image classification. In *2021 IEEECVF International Conference on Computer Vision ICCV* 3458–3468 (IEEE Computer Society, 2021).
15. Vu, Y. N. T. et al. MedAug: contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. In *Proc. 6th Machine Learning for Healthcare Conference* (eds Jung, K. et al.) 755–769 (PMLR, 2021).
16. Lu, M. Y., Chen, R. J. & Mahmood, F. Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding. In *Medical Imaging 2020: Digital Pathology* (eds. Tomaszewski, J. E. & Ward, A. D.) 11320J (SPIE, 2020).
17. Yang, P., Hong, Z., Yin, X., Zhu, C. & Jiang, R. Self-supervised visual representation learning for histopathological images. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (eds. de Bruijne, M. et al.) 47–57 (Springer International Publishing, 2021).
18. Srinidhi, C. L., Kim, S. W., Chen, F.-D. & Martel, A. L. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Med. Image Anal.* **75**, 102256 (2022).
19. DiPalma, J., Suriawinata, A. A., Tafe, L. J., Torresani, L. & Hassanpour, S. Resolution-based distillation for efficient histology image classification. *Artif. Intell. Med.* **119**, 102136 (2021).
20. Kiyasseh, D., Zhu, T. & Clifton, D. A. CLOCS: Contrastive Learning of Cardiac Signals across space, time, and patients. In *Proc. 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) 5606–5615 (PMLR, 2021).
21. Banville, H. J. et al. Self-supervised representation learning from electroencephalography signals. In *2019 IEEE 29th International Workshop on Machine Learning Signal Process MLSP* (IEEE Computer Society, 2019); https://doi.org/10.1109/MLSP.2019.8918693
22. Gopal, B. et al. 3KG: contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. in *Proc. Machine Learning for Health* (eds. Roy, S. et al.) 156–167 (PMLR, 2021).
23. Jiao, J. et al. Self-supervised contrastive video-speech representation learning for ultrasound. *Med. Image Comput. Comput. Assist. Interv.* **12263**, 534–543 (2020).
24. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
25. Xie, Y., Xu, Z., Zhang, J., Wang, Z. & Ji, S. Self-supervised learning of graph neural networks: a unified review. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022); https://doi.org/10.1109/TPAMI.2022.3170559
26. Meng, X., Ganoe, C. H., Sieberg, R. T., Cheung, Y. Y. & Hassanpour, S. Self-supervised contextual language representation of radiology reports to improve the identification of communication urgency. *AMIA Jt. Summits Transl. Sci. Proc.* **2020**, 413–421 (2020).
27. Girgis, H. Z., James, B. T. & Luczak, B. B. Identity: rapid alignment-free prediction of sequence alignment identity scores using self-supervised general linear models. *NAR Genom. Bioinform.* **3**, lqab001 (2021).
28. Li, Y. et al. BEHRT: transformer for electronic health records. *Sci. Rep.* **10**, 7155 (2020).
29. Wang, X., Xu, Z., Tam, L., Yang, D. & Xu, D. Self-supervised image-text pre-training with mixed data in chest X-rays. Preprint at https://doi.org/10.48550/arXiv.2103.16022 (2021).
30. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **4**, 86 (2021).
31. Li, F. et al. Fine-tuning Bidirectional Encoder Representations From Transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. *JMIR Med. Inform.* **7**, e14830 (2019).
32. Kraljevic, Z. et al. Multi-domain clinical natural language processing with MedCAT: the Medical Concept Annotation Toolkit. *Artif. Intell. Med.* **117**, 102083 (2021).
33. Kostas, D., Aroca-Ouellette, S. & Rudzicz, F. BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Front. Hum. Neurosci.* **15**, 653659 (2021).
34. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems* (eds Larochelle, H. et al.) 12449–12460 (Curran Associates, 2020).
35. Boyd, J. et al. Self-supervised representation learning using visual field expansion on digital pathology. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* 639–647 (IEEE Computer Society, 2021).

36. Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates, 2017).
37. Jaegle, A. et al. Perceiver IO: a general architecture for structured inputs and outputs. In *International Conference on Learning Representations* 4039 (ICLR, 2022).
38. Akbari, H. et al. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems* (eds. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) 24206–24221 (Curran Associates, 2021).
39. Nagrani, A. et al. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems* (eds Ranzato, M. et al.) 14200–14213 (Curran Associates, 2021).
40. Choromanski, K. et al. Masked language modeling for proteins via linearly scalable long-context transformers. Preprint at https://doi.org/10.48550/arXiv.2006.03555 (2020).
41. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
42. Rao, R. M. et al. MSA Transformer. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 8844–8856 (PMLR, 2021).
43. Lu, A. X., Zhang, H., Ghassemi, M. & Moses, A. Self-supervised contrastive learning of protein representations by mutual information maximization. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.04.283929 (2020).
44. Yang, C., Wu, Z., Zhou, B. & Lin, S. Instance localization for self-supervised detection pretraining. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 3986–3995 (IEEE Computer Society, 2021).
45. Jana, A. et al. Deep learning based NAS score and fibrosis stage prediction from CT and pathology data. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering BIBE* 981–986 (IEEE Computer Society, 2020).
46. Ohri, K. & Kumar, M. Review on self-supervised image recognition using deep neural networks. *Knowl. Based Syst.* **224**, 107090 (2021).
47. Holmberg, O. G. et al. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nat. Mach. Intell.* **2**, 719–726 (2020).
48. Spahr, A., Bozorgtabar, B. & Thiran, J.-P. Self-taught semi-supervised anomaly detection on upper limb X-rays. In *2021 IEEE 18th International Symposium on Biomedical Imaging ISBI* 1632–1636 (IEEE Computer Society, 2021).
49. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).
50. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
51. Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations* 1796 (ICLR, 2020).
52. Fedorov, A. et al. Tasting the cake: evaluating self-supervised generalization on out-of-distribution multimodal MRI data. Preprint at https://doi.org/10.48550/arXiv.2103.15914 (2021).
53. Li, Z. et al. Domain generalization for mammography detection via multi-style and multi-view contrastive learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (eds de Bruijne, M. et al.) 98–108 (Springer, 2021).
54. Endo, M., Krishnan, R., Krishna, V., Ng, A. Y. & Rajpurkar, P. Retrieval-based chest X-ray report generation using a pre-trained contrastive language-image model. in *Proc. Machine Learning for Health* (eds. Roy, S. et al.) 209–219 (PMLR, 2021).
55. Sriram, A. et al. COVID-19 prognosis via self-supervised representation learning and multi-image prediction. Preprint at https://doi.org/10.48550/arXiv.2101.04909 (2021).
56. Chen, R. J. & Krishnan, R. G. Self-supervised vision transformers learn visual concepts in histopathology. in *LMLR at Neural Information Processing Systems* (NeurIPS, 2021).
57. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
58. Logé, C. et al. Q-Pain: a question answering dataset to measure social bias in pain management. In *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks* (eds. Vanschoren, J. & Yeung, S.) 105 (NeurIPS, 2021).
59. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl Acad. Sci. USA* **117**, 12592 (2020).
60. Gamble, P. et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Commun. Med.* **1**, 14 (2021).

Acknowledgements

We thank A. Saporta and A. Tamkin for their helpful suggestions. We acknowledge support from the NIH grant UL1TR002550 to E.J.T.

Author contributions

All authors contributed to researching the literature and to the writing and editing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Pranav Rajpurkar.

Peer review information *Nature Biomedical Engineering* thanks Su-In Lee, Faisal Mahmood and Collin Stultz for their contribution to the peer review of this work.