2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)
Venice, Italy, April 8-11, 2019

# THE CONTINUOUS REGISTRATION CHALLENGE: EVALUATION-AS-A-SERVICE FOR MEDICAL IMAGE REGISTRATION ALGORITHMS

*K. Marstal*[⋆], *F. Berendsen*[†], *N. Dekker*[†], *M. Staring*[†], *S. Klein*[⋆]

[⋆] Biomedical Imaging Group Rotterdam (BIGR), Departments of Radiology & Medical Informatics,
Erasmus Medical Center, PO 2040, 3000 CA, Rotterdam, The Netherlands
k.marstal@erasmusmc.nl
[†] Division of Image Processing (LKEB), Department of Radiology,
Leiden University Medical Center, PO Box 9600, 2300 RC, Leiden, The Netherlands

## ABSTRACT

We have developed an open source, collaborative platform for researchers to develop, compare, and improve medical image registration algorithms. The platform handles data management, unit testing, and benchmarking of registration methods in a fully automatic fashion. In this paper we describe the platform and present the Continuous Registration Challenge. The challenge focuses on registration of lung CT and brain MR images and includes eight publicly available data sets. The platform is made available to the community as an open source project and can be used for organization of future challenges.

***Index Terms*—** image registration; grand challenges; reproducibility

## 1. INTRODUCTION

Medical image registration is the process of transforming images into a common coordinate system such that corresponding pixels represent homologous biological points. This is a common preprocessing step in many applications including segmentation of anatomical structures and computer-aided diagnosis. Each application requires the end-user to carefully select a registration method and tune its hyper-parameters. However, registration methods are implemented in many different toolboxes each with their own user interface, conventions, and input and output formats. This lack of standardization makes image registration methods difficult to compare.

To standardize comparison, registration methods can be evaluated in a Grand Challenge (GC). GCs are competitions with standardized data sets, evaluation methods, and experimental setups that focus on specific research topics. The experiments are run by third parties which ensures fair, independent evaluations. In the field of image registration, initiatives such as Evaluation of Methods for Pulmonary Image Registration 2010 (EMPIRE10) [1], Non-Rigid Image Registration Evaluation Project (NIREP) [2], and the Retrospective Image Registration Evaluation (`http://www.insight-journal.org/rire`) have aimed at standardizing evaluation of medical image registration algorithms. These initiatives have provided valuable insights to the community.

However, many GCs suffer from a number of limitations. Firstly, GCs are often static, one-time events that use closed source evaluation systems. This prevents new methods and new data sets from being included in the challenge. Secondly, participants are usually given fixed a test data set which introduces risk of over-tuning and leaderboard climbing. Thirdly, participants submit processed data, not code, to the challenge. Without access to source code it is practically impossible for other researchers to independently verify results or apply the methods to their own data. These limitations conflict with the scientific principle of reproducibility.

To address these limitations, and inspired by modern software development practices, we developed a platform for running GCs that we named SuperBench. The platform manages data, code compilation, registration, evaluation, visualization, and leaderboard generation and random subsets of data. The platform was designed to repeatedly run large experiments similar in spirit to the Continuous Integration testing methodology [3]. New methods and data sets can be added to SuperBench by checking code into an open source repository. The repository contains all information for running experiments both on large clusters and local machines. Researchers can therefore easily verify results.

Using this system, we launched the Continuous Registration Challenge (CRC) for lung and brain images. CRC is a collaborative challenge where researchers develop, compare, and continuously improve algorithms and parameter settings. We implemented a widely used registration method for baseline comparison and included eight pubicly available data sets in the challenge. CRC is now open for submissions and we plan to keep it open in the coming years. Registrations are run nightly and results are available at `https://continuousregistration.grand-challenge.org`.

978-1-5386-3640-4/19/$31.00 ©2019 IEEE

# 2. METHODS

## 2.1. The SuperBench Framework

SuperBench consists of the ”SuperElastix” C++ framework for running registrations [4] and a set of tools for orchestrating experiments. It is designed as a Continuous Integration system that runs all methods on all data sets and publishes results online. The architecture is shown in Figure 1.

To add a registration method to the system, participants implement a ”component” in SuperElastix and provide a ”parameter file” in JSON or XML format that describes how SuperElastix should run that component. The parameter file defines all parameter settings that the component requires. The component and parameter file are submitted to the online code repository. To add data to a challenge, organizers make the data set available online and implement functionality in SuperBench that fetches data online or loads data from disk after manual download. To run an experiment, the repository is downloaded, compiled, unit tested, and run on a cluster or on a local machine. The entire process can be scripted. Registrations output deformation fields that are used to evaluate ground truth. Evaluation results are outputted as JSON files for subsequent processing and HTML tables for online publishing. The entire pipeline is run via Jenkins (`https://jenkins.io`). Compilation logs are made available online via CDash [5].

This design was chosen to fulfill three goals: Firstly, we aim to provide Evaluation-as-a-Service (EaaS) [6] so researchers can focus more on developing registration methods and less on infrastructure for running experiments. Secondly, participants must submit methods to the online code repository to ensure that methods were always available to third parties. Lastly, we want to promote collaborative coding and have researchers share and improve on each others work. For example, if one participant implementes a good initialization strategy, other participants should be able to take advantage of that approach for their own method.

SuperElastix [4] is responsible for running individual registrations. Briefly, SuperElastix is a modular image registration toolbox written in C++ that provides standard interfaces for reading and writing images and configuring registration methods. Internally, SuperElastix consists of a set of registration components and a core framework that connects and executes components. Components are isolated, modular code blocks that send and receive data via a predefined set of interfaces. A component can be a complete registration algorithm or subcomponent thereof. The core framework consists of a thin layer that is responsible for instantiating components, connecting components in a network as described by parameter file, and executing the components in the right order. When a network is run, SuperElastix collects citation information from all components and displays it to the user.

At runtime SuperElastix executes the network similar to how TensorFlow [7], CNTK [8], and Caffe2 [9] execute computational graphs for deep learning. However, in addition to exchange of data buffers, connections in SuperElastix can be control flow, function calls, and online adjustments of component settings.

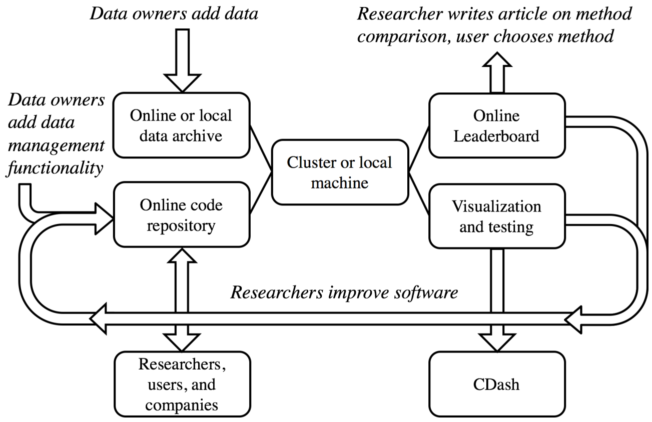
Data owners add data
Researcher writes article on method comparison, user chooses method
Data owners add data management functionality
Online or local data archive
Online code repository
Cluster or local machine
Online Leaderboard
Visualization and testing
Researchers improve software
Researchers, users, and companies
CDash

**Fig. 1**. Architecture. The framework is fully automated and provides infrastructure for challenge organizers, a feedback-loop for algorithm developers, and a ready-to-use registration library for end-users.

## 2.2. The Continuous Registration Challenge

The challenge focuses on pairwise registration of lung images and brain images.

**Data** We included five public brain data sets, two public lung data sets, and one private lung data set from our own institution. We included the ISBR18, LPBA40, CUMC12, and MGH10 brain MR data sets that were previously used in [10] for a comprehensive evaluation on 14 registration methods. We also included the HAMMERS brain MR data set [11]. For lungs, we included the widely used POPI [12] and DIR-LAB [13] data sets. The SPREAD [14] private lung data set was included to discourage overfitting to the public lung data sets. The data sets are listed in Table 1.

**Evaluation** Point sets are evaluated using Target Registration Error (TRE) and Hausdorff. Segmentations are evaluated using Dice Similarity Coefficient (DSC), Union Coefficient (UC), Jaccard Coefficient (JC), False Negative Error (FNE), False Positive Error (FPE), and Volume Similarity (VS).

**Rules** Participants can submit multiple registration methods and multiple parameter files with different registration settings. Participants can also submit parameter files with for existing components. We provide compute hours at a cluster at our own institution. Participants have a maximum of one hour of wall clock time per registration of a pair of images. Partcipants can use external data (e.g. a pretrained TensorFlow model) and have the build script automatically download it at build time.

**Registration methods** As a baseline example, we integrated Elastix [15] as a component into SuperElastix, and

| Name | Availability | Modality | # Subjects | Anatomy | Type | Ground truth |
|---|---|---|---|---|---|---|
| **POPI** | Public | CT | 6 | Lung | Intra-subject | Points |
| **DIR-LAB** | Public | CT | 10 | Lung | Intra-subject | Points |
| **SPREAD** | Private | CT | 21 | Lung | Intra-subject | Points |
| **LPBA40** | Public | MR | 40 | Brain | Inter-subject | Segmentations |
| **ISBR18** | Public | MR | 18 | Brain | Inter-subject | Segmentations |
| **CUMC12** | Public | MR | 12 | Brain | Inter-subject | Segmentations |
| **MGH10** | Public | MR | 10 | Brain | Inter-subject | Segmentations |
| **HAMMERS** | Public | MR | 20 | Brain | Inter-subject | Segmentations |

**Table 1**. Data sets.

submitted three methods to the challenge: 1) no registration (*Identity*); 2) affine registration (*Elastix Affine*); and 3) nonrigid B-spline registration (*Elastix Affine+BSpline*). We use Mutual Information (MI), three levels of resolution, and Adaptive Stochastic Gradient Descent (ASGD). We used a B-spline grid spacing of $12 \times 12 \times 12$ mm for lungs and $4 \times 4 \times 4$ mm for brains, but otherwise we did not tune elastix for specific data sets. We also implemented NiftyReg [16], ITKv4 SyN [17], and ITKv4 ANTs [18], but do not show results since these methods needed further tuning at the time of writing.

## 3. RESULTS

The results are listed in Table 2. We only report TRE and DSC for brevity. CUMC12 and LPBA40 registration results are comparable to that of [10]. CUMC12 and ISBR18 registrations seemed to have failed because of a B-spline grid that was too dense coupled with the fact that no regularization was used. The POPI and DIR-LAB results are almost on par with previous work. We did not tune the elastix parameters to any particular data set. The mixed results show the importance of applying a registration method to a wide variety of data sets. The full result table is available at the challenge website.

## 4. DISCUSSION

We developed a platform for evaluation of medical image registration algorithms. The goal was to design a system for grand challenges that provides automated and standardized benchmarking, collaboration on code level, and reproducible experiments. We adopted a test driven development methodology to help participants produce code of higher clarity, readability, and robustness. The development model encourages collaboration in contrast to most grand challenges which are based solely on competition. We hope that this approach will foster progression in the field, simplify adoption of new registration methods, and harmonize reporting of results in scientific articles.

The SuperBench framework brings best practices from modern software development to organizers of grand challenges in image registration. This includes a robust build system, a unit testing framework, and a continuous integration system. SuperElastix is able to accommodate multiple registration paradigms such as methods based on b-splines, diffeomorphic registration, and velocity fields. In this work, we focused on the SuperBench framework and the launch of CRC. We will write a follow-up paper and present results when the challenge has more submissions.

Our design choices come with several trade-offs. Firstly, we adopt an open source approach which requires participants to submit code during development. This ensures that code is always available to third parties, but many researchers may prefer to keep their code private at least until their method has been published in a journal. Secondly, to automate execution and evaluation of registration algorithms, we dictate an API that users have to follow. The associated learning curve may be an entry barrier to some participants. Finally, we include public data sets to ensure results can be reproduced but this prevents us from holding out a hidden data set for evaluation of methods. Particants therefore have access to all data including the subset that is used to evaluate results for the leaderboard. However, if methods are over-fitted to a particular data set, at least it is fully transparent to which data set a method has been tuned.

## 5. CONCLUSION

We developed an automated platform for running grand challenges that is made available to the community as an open source project. The platform ensures that results can be independently verified, that methods can be easily applied to new data sets, and that methodological improvements can be benchmarked against existing algorithms.

## 6. ACKNOWLEDGEMENTS

This research was supported by NWO-TTW, project 13351, Medical Image Registration: Linking Algorithm and User.

| Name | Identity | Elastix Affine | Elastix Affine+BSpline |
|---|---|---|---|
| **POPI (TRE)** [12] | $8.09 \pm 2.73$ | $8.03 \pm 2.91$ | $1.58 \pm 0.59$ |
| **DIR-LAB (TRE)** [13] | $8.46 \pm 3.16$ | $8.35 \pm 3.53$ | $2.91 \pm 1.80$ |
| **SPREAD (TRE)** [14] | $493.44 \pm 471.23$ | $3.96 \pm 2.74$ | $2.05 \pm 1.74$ |
| **LPBA40 (DSC)** [10] | $0.61 \pm 0.03$ | $0.69 \pm 0.01$ | $0.73 \pm 0.02$ |
| **ISBR18 (DSC)** [10] | $0.21 \pm 0.12$ | $0.38 \pm 0.03$ | $0.35 \pm 0.02$ |
| **CUMC12 (DSC)** [10] | $0.45 \pm 0.05$ | $0.50 \pm 0.04$ | $0.65 \pm 0.03$ |
| **MGH10 (DSC)** [10] | $0.54 \pm 0.03$ | $0.55 \pm 0.03$ | $0.51 \pm 0.03$ |
| **HAMMERS (DSC)** [11] | $0.54 \pm 0.06$ | $0.69 \pm 0.02$ | $0.73 \pm 0.03$ |

**Table 2**. Results.

## 7. REFERENCES

[1] K. Murphy, B. Van Ginneken, J. Reinhardt, et al., "Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge," *IEEE Transactions on Medical Imaging*, vol. 30, no. 11, pp. 1901–1920, 2011.

[2] G. Christensen, X. Geng, J. Kuhl, et al., "Introduction to the non-rigid image registration evaluation project (NIREP)," in *International workshop on biomedical image registration*. Springer, 2006, pp. 128–135.

[3] S. Stolberg, "Enabling agile testing through continuous integration," in *Agile Conference, 2009. AGILE'09.* IEEE, 2009, pp. 369–374.

[4] F. Berendsen, K. Marstal, S. Klein, and M. Staring, "The design of SuperElastix–a unifying framework for a wide range of image registration methodologies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 58–66.

[5] Ken Martin and Bill Hoffman, *Mastering CMake: a cross-platform build system*, Kitware, 2010.

[6] A. Hanbury, Henning H. Müller, K. Balog, et al., "Evaluation-as-a-service: Overview and outlook," *arXiv preprint arXiv:1512.07454*, 2015.

[7] M. Abadi, P. Barham, J. Chen, et al., "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[8] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2135–2135.

[9] Y. Jia, E. Shelhamer, J. Donahue, et al., "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[10] A. Klein, J. Andersson, B. Ardekani, et al., "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[11] I. Gousias, A. Edwards, M. Rutherford, et al., "Magnetic resonance imaging of the newborn brain: Manual segmentation of labelled atlases in term-born and preterm infants," *Neuroimage*, vol. 62, no. 3, pp. 1499–1509, 2012.

[12] J. Vandemeulebroucke, S. Rit, J. Kybic, et al., "Spatiotemporal motion estimation for respiratory-correlated imaging of the lungs," *Medical physics*, vol. 38, no. 1, pp. 166–178, 2011.

[13] R. Castillo, E. Castillo, R. Guerra, et al., "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets," *Physics in Medicine & Biology*, vol. 54, no. 7, pp. 1849, 2009.

[14] J. Stolk, H. Putter, E. Bakker, et al., "Progression parameters for emphysema: A clinical investigation," *Respiratory medicine*, vol. 101, no. 9, pp. 1924–1930, 2007.

[15] S. Klein, M. Staring, K. Murphy, et al., "Elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[16] Marc Modat, *Efficient dense non-rigid registration using the free-form deformation framework*, Ph.D. thesis, UCL (University College London), 2012.

[17] B. Avants, M. Grossman, J. Gee, et al., "Symmetric diffeomorphic image registration: Evaluating automated labeling of elderly and neurodegenerative cortex and frontal lobe," in *International Workshop on Biomedical Image Registration*. Springer, 2006, pp. 50–57.

[18] B. Avants, N. Tustison, G., Song, et al., "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.