# Adrian Vasile Dalca — Research Statement

My research focuses on machine learning techniques and probabilistic models for healthcare, with an emphasis on the analysis of medical images. As the availability of biomedical data explodes, we have an unprecedented opportunity to accelerate our understanding of fundamental scientific concepts and improve clinical practice. In my research, I seize this opportunity by developing systems that learn to extract knowledge from complex biomedical data, starting with medical images. My goal is to advance techniques in machine learning and computer science that lead to improved understanding of disease and patient care. My main contributions have been published in the top technical conferences and journal in medical image analysis (MICCAI, IPMI, IEEE TMI), computer vision (CVPR), and machine learning (NIPS).

In the past, I developed new methods and systems to process heterogeneous imaging and health data quickly, extract meaningful representations, and transform them into usable information for scientific study. In the future, I will develop new frameworks for learning patient models that integrate signals across different biomedical data domains. Since expert annotations, or accurate labels, remain sparse, I will develop unsupervised, weakly supervised, and transfer learning methods to enable previously unattainable analysis across diverse datasets. I will use these analytic methods to establish improved clinical workflows, where new prediction and treatment techniques exploit the richness of the data.

My approach to research stems from my excitement for solving technical challenges with clinical impact. I strive to first gain an understanding of the clinical or scientific application, then develop analytic solutions. I focus on medical image analysis, a challenging field that plays an increasingly important role in healthcare, and also provides an avenue to advance computer science techniques more broadly. Consequently, I contribute to and draw from a range of technical fields, including machine learning, medical image analysis, and computer vision. I strive to develop intuitive explanations that connect analytic techniques with real world interpretations. Several of my open-source solutions are actively downloaded and are already being used in large clinical analyses to help analyze population-wide correlations of image features, genetic variants and clinical markers.

## Current Research

Millions of diagnostic medical records are acquired every day as part of standard medical practice, with biomedical scans playing an important role. There is a wealth of knowledge to be gained from analyzing clinically-sourced images, which will lead to better understanding of disease and treatments for patients. Unfortunately, clinical scans come with significant heterogeneity, noise, and sparsity because of the variety of medical centers, acquisition technologies, and patient populations. This is in stark contrast to the small, clean, curated and labelled datasets frequently used for method development, which are often far removed from applicability to real populations. The analyses of clinical images is an important and challenging area of research, providing an opportunity for the development of technical methods that can have clinical impact, as well as wide applicability even beyond imaging.

Several of my methods are motivated by a clinical stroke study - a long-standing collaboration that aims to identify genetic underpinnings of cerebrovascular disease by learning from thousands of medical scans and clinical markers. The heterogeneous data come from twelve hospitals around the world.

### Inference in clinical quality data

Three-dimensional magetic resonance (MR) scans are acquired as stacks of 2D slices. In the majority of clinical settings, such as the stroke cohort, scans can contain very sparse, noisy slices, having less than 15% of the data found in densely-stacked "research quality" scans. To facilitate the use of clinical scans for scientific study, I defined *medical image imputation* – the statistical inference of unobserved anatomical slices using only a collection of low resolution volumes [1, 2]. I specified a probabilistic mixture model that explicitly accounts for missing data across a population, and showed that missing anatomy can be estimated using expectation maximization. The method exploits the fact that each low-resolution scan captures some sparse aspect of local fine-scale anatomical structure intrinsically shared in a population. The resulting restored scans enable subsequent analysis not previously possible in a broad range of problems.

Imputed clinical scans can be used to analyse cerebrovascular pathologies and understand cerebral ischemia. I tackled the challenge of simultaneously identifying different spatially varying diseases that present with similar features on MR scans. Leveraging clinicians' rules and intuitions used in practice, I developed a probabilistic model that learned spatial disease priors from a limited collection of scans to capture anatomical and disease variability [3]. I used this anatomical representation to describe patterns of disease spread and growth, facilitating long-term health predictions for each specific patient. I developed a mathematical model of intensity, shape, and spatial distribution of different pathologies, and an algorithm that successfully annotates clinical brain scans in stroke patients, separately identifying stroke and other white matter disease.

## Efficient large scale analyses

Aligning, or registering, medical images is a central task for clinical population analyses. Traditional registration methods optimize an objective independently for each pair of scans, requiring tens of minutes to hours per scan pair – prohibitively slow for large analyses. During my post doc, I worked with graduate students to develop a new type of registration framework, called VoxelMorph, that learns a parametrized registration function using a collection of volumes, building image representations using convolutional neural networks [4, 5]. The method enables registration of two MRI volumes in under a second, and is successful even in small training regimes. By improving image registration runtime by orders of magnitude, we can align thousands of scans from a study in minutes.

Registration of *clinical* images is particularly challenging because of their low resolution. I previously devised an algorithm that models sparsity to improve registration, leading to a best paper award [6]. Recently, we combined these ideas with VoxelMorph, leading to fast and accurate registration of low quality clinical scans [7].

I drove the development of several of these papers, and others were led by graduate students whom I supervised. Our publications on registration have resulted in best paper awards, hundreds of code downloads, and active use in large studies.

## Biomedical data and anatomy

Imaging genetics is a nascent field that uses imaging as a rich quantitative phenotype to study genetic mechanisms underlying disease. Most imaging genetics methods assume simple relationships between anatomical features and genotypes, often ignoring the structure and relationships across domains. My colleagues and I developed a joint probabilistic model for images, genotypes and neurological disease diagnosis that enables a richer characterization of their relationship [8]. The method improves detection sensitivity and specificity, yielding relevant features and variants that were not otherwise detectable.

Despite a growing body of work using imaging for genetic discovery, there has been limited modeling of genetic effects on anatomy. I explored this direction with a simple question: can genetic variants predict aspects of MR scans? I developed a generative Gaussian Process model for patient-specific anatomical predictions over time, using external variables [9]. I used this model to predict, MRI scans exhibiting neurodegeneration or white matter disease spread in dementia patients. The model captures anatomical change through a combination of population-wide regression and a description of the subject's health using genetic and clinical indicators. I demonstrated that these health factors have a significant impact on the subject's anatomy. My colleageus and I have recently built on these ideas, and developed a model that enables improved sensitivity in genetic variant detection and accurate phenotype predictions [10]. The method jointly learns subject representations using a Gaussian Process prior based on genetic and clinical signals with convolutional neural networks-based characterizations of images.

Despite its importance, imaging genetics lacked a strong presence in the medical imaging community. I founded two workshops at MICCAI, the main medical image analysis conference, *MICGen: Workshop on Imaging Genetics*, and *BeyondMIC: Integrating Imaging and non-Imaging Modalities for Healthcare Challenges*. We have organized these workshops for four years, fostering a thriving interdesciplinary culture and growing attendance, and resulting in several proceedings and an edited book [11]. I look forward to continuing to expand this effort in the future.

# Research Plans

I am passionate about developing new methods that create a holistic view of a subject's current and future health states using a diverse set of biomedical data sources.

## Biomedical representations to characterize variability

As data heterogeneity increases, an important goal is to distinguish between meaningful multi-modal variation that we want to characterize and use, and nuisance factors, such as variations in protocols, acquisition technology, medical definitions, and clinical conventions, that we want to eliminate. To enable a more personalized view of patient health, I plan to develop machine learning frameworks to tease out the many sources of biomedical variability and their interactions.

I am particularly interested in using multi-modal biomedical data to reduce uncertainty of predictions. Even in well-defined tasks with curated data, such as medical image segmentation, we often aim to reproduce a *ground truth* target, despite frequent disagreement among clinical experts about what the ground truth is. Patient data external to a medical scan can help resolve this disagreement, lowering inconsistency. Furthermore, probabilistic anatomical models provide a natural way to estimate uncertainty [12]. I aim to build probabilistic patient health representations that explicitly model interactions between such external biomedical factors and images. I will use my experience with the multi-site stroke cohort to develop mechanisms that counter and consistently eliminate unwanted biases.

Clinical abnormality is often measured in terms of deviations from the population norm despite high levels of inter-patient variability. I will build on our recent work that jointly models population structure and anatomical features [10] to devise new methods that identify subgroup-specific longitudinal trends, biomedical factors, and environments. Using statistical analyses of resulting characterizations, I plan to develop a new definition of what is abnormal for a specific patient to better inform treatment decisions.

## Large data with limited annotations

Despite increasing availability of raw data, most of it lacks manual annotations. Building on my recent work on unsupervised learning [4, 12], I will design methods for transfer learning, weakly supervised learning and zero-shot learning that exploit population-wide structure to create meaningful clinical annotations.

I am particularly interested in learning representations that characterize both available and missing data across multiple medical domains. These models should capture only information relevant to diagnoses and clinical annotations, while discarding information not necessary for the task. Compact representations require far fewer annotations for accurate prediction. We have already shown that a learned minimal description of anatomical shape can enable anatomical delineation in a new collection of unannotated scans [12].

Building on my work on image imputation, I am eager to develop general imputation methods for healthcare data. To tackle this, I will leverage correlations of relevant information across multiple domains and timepoints to build better estimates for any missing data, with the goal of improving healthcare predictions.

## Improved medical workflows

Current clinical practice involves summarizing measurements into a medical report. Radiologists summarize medical scans using specialized vocabulary, and lab values are often binned: if the blood pressure passes a threshold, the patient is hypertensive. Summaries enable a smooth medical workflow, at the cost of discarding rich signals. By building extensive characterizations of biomedical data and records, I aim to redefine this pipeline to leverage the entire richness of medical images or clinical measures to inform treatment.

My goal is to establish a computational mechanism to learn and use clinical features that surpass rigid human rules. For example, I plan to build methods that automatically learn anatomical features relevant to a medical diagnosis or genetic allele, which in turn can provide new insights into mechanisms of pathogenesis. Similarly, longitudinal traits are hard to define and evaluate manually, while it should be possible to learn complex longitudinal features to characterize patient measurements over time.

I envision an initial set of systems to automatically translate medical measurements into clinical reports. I believe we can learn features that will be *more* effective than current summaries at highlighting important aspects for a given task. These systems will not only recommend improved treatment and track outcomes, but predict possible side effects, describe patient-specific disease progression, and adapt therapy as necessary.

## Conclusion

I believe that successful analytic solutions that impact clinical research and treatment are rooted in close clinical collaborations and draw on technical insights from computer science, signal processing, and statistics. In my research, I developed and published several machine learning methods and probabilistic models that enabled analysis of thousands of patients. In turn, these methods facilitated clinical findings that have been presented in several clinical papers, talks, and posters.

I thoroughly enjoy working with students and have supervised several undergraduate and graduate students. I draw inspiration from multiple disciplines and enjoy actively collaborating on projects in machine learning, computer vision and computational imaging. Outside the field of medical imaging, my colleagues and I developed a method to reconstruct video of inter-stellar objects [13], demonstrated a way to synthesize humans based on new poses [14], and developed tipiX [15] - a rapid image visualization tool that enables the exploration of large vision datasets.[1] These methods touch a wide range of different fields, yet have direct applicability to medical imaging and healthcare. For example, tipiX is now an interactive rapid visualization method for high dimensional medical image data, and received a best paper award for impact and usability.

I will contribute to technical advancements in a breath of fields, with a focus on machine learning for healthcare applications. I am excited to collaborate on natural language processing to connect biomedical data with clinical narratives, computer graphics to improve medical image synthesis and data imputation methods, and video processing to analyze and predict temporal and longitudinal biomedical data.

## References

(*) senior role


1. **Adrian V. Dalca**, Katherine Bouman, William Freeman, Mert Sabuncu, Natalia Rost, and Polina Golland
"Population Based Image Imputation"
*IPMI: Information Processing in Medical Imaging* 10265 (2017), 659–671.
**Best poster award.**

2. **Adrian V. Dalca**, Katherine Bouman, William Freeman, Mert Sabuncu, Natalia Rost, and Polina Golland
"Medical Image Imputation from Image Collections"
*IEEE TMI: Transactions on Medical Imaging* (2018), In Press.

3. **Adrian V. Dalca**, Ramesh Sridharan, Lisa Cloonan, Kaitlin M. Fitzpatrick, Allison Kanakis, Karen L. Furie, Jonathan Rosand, Ona Wu, Mert Sabuncu, Natalia S. Rost, and Polina Golland
"Segmentation of Cerebrovascular Pathologies in Stroke Patients with Spatial and Shape Priors"
*MICCAI: Medical Image Computing and Computer Assisted Intervention, LNCS* 8674 (2014), pp. 773–780.

4. **Adrian V. Dalca**, Guha Balakrishnan, John Guttag, and Mert Sabuncu
"Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration"
*MICCAI: Medical Image Computing and Computer Assisted Intervention, LNCS* 11070 (2018), 729–738.
**Oral Presentation. Finalist for Best Paper (Young Scientist) award.**

5. Guha Balakrishnan, Amy Zhao, Mert Sabuncu, John Guttag, and **Adrian V. Dalca***
"An Unsupervised Learning Model for Deformable Medical Image Registration"
*CVPR: Computer Vision and Pattern Recognition* (2018).

6. **Adrian V. Dalca**, Andreea Bobu, Natalia S Rost, and Polina Golland
"Patch-Based Discrete Registration of Clinical Brain Images"
*MICCAI PATCHMI: Patch-based Techniques in Medical Imaging, LNCS* 9993 (2016), 60–67.
**Oral Presentation. Best paper award.**

7. Kathleen Lewis, Balakrishnan Guha, Natalia Rost, John Guttag, and **Adrian V. Dalca***
"Registration of Sparse Clinical Images"
*NIPS ML4H: Machine Learning for Health* (2018),
**Spotlight**.


[1] TipiX was originally devised to visualize my Boston Timescape project – a collection of more than one million curated pictures of the Boston skyline that I have taken over seven years. The collection is available at `http://bostontimescape.com`.

8. Nematollah K. Batmanghelich, **Adrian V. Dalca**, Gerald Quon, Mert R. Sabuncu, and Polina Golland
"Probabilistic Modeling of Imaging, Genetics and Diagnosis"
*IEEE TMI: Transactions on Medical Imaging* 35.7 (2016), pp. 1765–79.

9. **Adrian V. Dalca**, Ramesh Sridharan, Mert Sabuncu, and Polina Golland
"Predictive Modeling of Anatomy with Genetic and Clinical Data"
*MICCAI: Medical Image Computing and Computer Assisted Intervention, LNCS* 9351 (2015), pp. 519–526.

10. Francesco P. Casale, **Adrian V. Dalca**, Luca Saglietti, Nicolo Fusi, and Jennifer Listgarten
"Gaussian Process Prior Variational Autoencoders"
*NIPS: Neural Information Processing Systems* (2018).

11. **A. V. Dalca**, Kayhan N Batmanghelich, Mert R Sabuncu, and Li Shen (Eds)
*Imaging Genetics*
Elsevier, 2017.

12. **Adrian V. Dalca**, John Guttag, and Mert Sabuncu
"Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation"
*CVPR: Computer Vision and Pattern Recognition* (2018).

13. Katherine L. Bouman, Michael D. Johnson, **Adrian V. Dalca**, Andrew A. Chael, Freek Roelofs, Sheperd S. Doeleman, and William T. Freeman
"Reconstructing Video of Time-Varying Sources from Radio Interferometric Measurements"
*IEEE TCI: Transactions on Computational Imaging* 4.4 (2018), pp. 512–527.

14. Guha Balakrishnan, Amy Zhao, **Adrian V. Dalca**, Fredo Durand, and John Guttag
"Synthesizing Images of Humans in Unseen Poses"
*CVPR: Computer Vision and Pattern Recognition* (2018),
**Oral Presentation**.

15. **Adrian V. Dalca**, Ramesh Sridharan, Natalia S. Rost, and Polina Golland
"tipiX: Rapid Visualization of Large Image Collections"
*MICCAI IMIC: Interactive Medical Image Computing Workshop* (2014),
**Oral Presentation. Best paper award for impact and usability.**