AI and machine learning in biostatistical research: example projects and IT requirements

Valeria Vitelli
Oslo Center for Biostatistics and Epidemiology (OCBE)
Department of Biostatistics, University of Oslo

September 6, 2018

Outline

Oslo Center for Biostatistics and Epidemiology (OCBE)

- OCBE is a joint center integrating the activities of the Department of Biostatistics, UiO and the Section of Biostatistics, Epidemiology and Health Economics, OUS;
- people at OCBE deal with methodological research in all areas of biostatistics & machine learning;

Oslo Centre for Biostatistics and Epidemiology (OCBE)

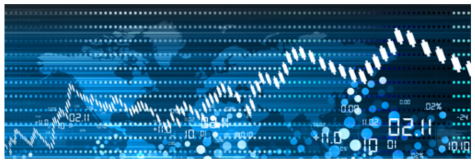


Photo illustration: Colourbox.no

Oslo Centre for Biostatistics and Epidemiology (OCBE) is a joint center integrating the activities of the Department of Biostatistics, UiO and the Section of Biostatistics, Epidemiology and Health Economics, OUS

A joint centre with

Oslo University Hospital

With funding from

HELSE SØR-ØST

Contact

Postal address
P.O.Box 1122 Blindern

Oslo Center for Biostatistics and Epidemiology (OCBE)

- OCBE is a joint center integrating the activities of the Department of Biostatistics, UiO and the Section of Biostatistics, Epidemiology and Health Economics, OUS;
- people at OCBE deal with methodological research in all areas of biostatistics & machine learning;

Oslo Centre for Biostatistics and Epidemiology (OCBE)

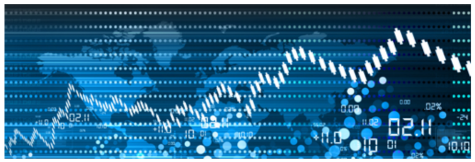


Photo illustration: Colourbox.no

Oslo Centre for Biostatistics and Epidemiology (OCBE) is a joint center integrating the activities of the Department of Biostatistics, UiO and the Section of Biostatistics, Epidemiology and Health Economics, OUS

A joint centre with

Oslo University Hospital

With funding from

HELSE SØR-ØST

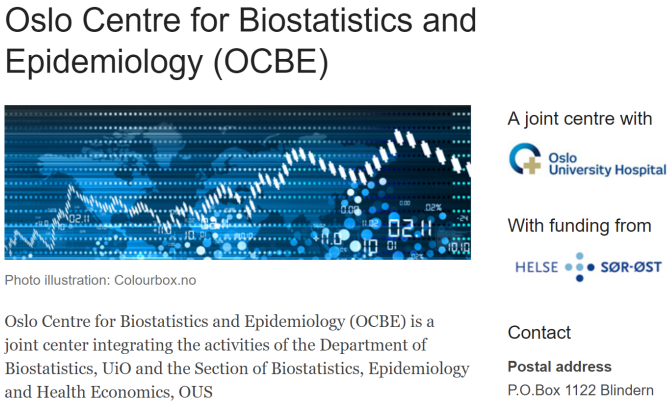
Contact

Postal address
P.O.Box 1122 Blindern

- OCBE provides statistical and epidemiological training for researchers and students of the Medical Faculty of the University of Oslo, of the Oslo University Hospital and of Helse Sør-Øst (HSØ);

Oslo Center for Biostatistics and Epidemiology (OCBE)

- OCBE is a joint center integrating the activities of the Department of Biostatistics, UiO and the Section of Biostatistics, Epidemiology and Health Economics, OUS;
- people at OCBE deal with methodological research in all areas of biostatistics & machine learning;



- OCBE provides statistical and epidemiological training for researchers and students of the Medical Faculty of the University of Oslo, of the Oslo University Hospital and of Helse Sør-Øst (HSØ);
- the advising activity at OCBE is directed towards all areas of medicine and health related research, from clinical and epidemiological research, to molecular biology and other basic medical sciences.

OCBE in a nutshell: **research groups**

- **Stochastic models and inference:** innovative models for representing biological principles / patterns of dependence, computationally intensive inference in the life sciences, statistical genomics, big data.

OCBE in a nutshell: **research groups**

- **Stochastic models and inference:** innovative models for representing biological principles / patterns of dependence, computationally intensive inference in the life sciences, statistical genomics, big data.
- **Statistical learning in molecular medicine:** development, validation and application of statistical machine learning methods for clinically relevant predictions based on high-dimensional data from molecular data sources.

OCBE in a nutshell: **research groups**

- **Stochastic models and inference:** innovative models for representing biological principles / patterns of dependence, computationally intensive inference in the life sciences, statistical genomics, big data.
- **Statistical learning in molecular medicine:** development, validation and application of statistical machine learning methods for clinically relevant predictions based on high-dimensional data from molecular data sources.
- **Measurement error modeling:** understanding the behavior of standard statistical methods in non-standard settings, understand the effect of the error in -omics measurements and exposure / response measurements.

OCBE in a nutshell: **research groups**

- **Stochastic models and inference:** innovative models for representing biological principles / patterns of dependence, computationally intensive inference in the life sciences, statistical genomics, big data.
- **Statistical learning in molecular medicine:** development, validation and application of statistical machine learning methods for clinically relevant predictions based on high-dimensional data from molecular data sources.
- **Measurement error modeling:** understanding the behavior of standard statistical methods in non-standard settings, understand the effect of the error in -omics measurements and exposure / response measurements.
- **Infectious diseases:** math modeling of infectious diseases, stat learning for network data, structure and dynamical properties of social networks.

OCBE in a nutshell: **research groups**

- **Stochastic models and inference:** innovative models for representing biological principles / patterns of dependence, computationally intensive inference in the life sciences, statistical genomics, big data.
- **Statistical learning in molecular medicine:** development, validation and application of statistical machine learning methods for clinically relevant predictions based on high-dimensional data from molecular data sources.
- **Measurement error modeling:** understanding the behavior of standard statistical methods in non-standard settings, understand the effect of the error in -omics measurements and exposure / response measurements.
- **Infectious diseases:** math modeling of infectious diseases, stat learning for network data, structure and dynamical properties of social networks.
- **Epidemiological studies of lifestyle and chronic diseases:** missing data and bias in exposure-disease associations in epidemiological studies, lifestyle and risk of chronic diseases.

OCBE in a nutshell: **research groups**

- **Stochastic models and inference:** innovative models for representing biological principles / patterns of dependence, computationally intensive inference in the life sciences, statistical genomics, big data.
- **Statistical learning in molecular medicine:** development, validation and application of statistical machine learning methods for clinically relevant predictions based on high-dimensional data from molecular data sources.
- **Measurement error modeling:** understanding the behavior of standard statistical methods in non-standard settings, understand the effect of the error in -omics measurements and exposure / response measurements.
- **Infectious diseases:** math modeling of infectious diseases, stat learning for network data, structure and dynamical properties of social networks.
- **Epidemiological studies of lifestyle and chronic diseases:** missing data and bias in exposure-disease associations in epidemiological studies, lifestyle and risk of chronic diseases.
- **Causal inference methods:** analyzing the causal conclusions that can be drawn from statistical data, also for longitudinal and time to event data.

OCBE in a nutshell: **research groups**

- **Stochastic models and inference:** innovative models for representing biological principles / patterns of dependence, computationally intensive inference in the life sciences, statistical genomics, big data.
- **Statistical learning in molecular medicine:** development, validation and application of statistical machine learning methods for clinically relevant predictions based on high-dimensional data from molecular data sources.
- **Measurement error modeling:** understanding the behavior of standard statistical methods in non-standard settings, understand the effect of the error in -omics measurements and exposure / response measurements.
- **Infectious diseases:** math modeling of infectious diseases, stat learning for network data, structure and dynamical properties of social networks.
- **Epidemiological studies of lifestyle and chronic diseases:** missing data and bias in exposure-disease associations in epidemiological studies, lifestyle and risk of chronic diseases.
- **Causal inference methods:** analyzing the causal conclusions that can be drawn from statistical data, also for longitudinal and time to event data.
- **Probabilistic Inference Lab (PIL):** more later....

OCBE in a nutshell: **teaching**

- **PhD courses.** Introduction to infectious disease modelling; Quantitative biology, or mathematics is biology's next microscope; Prediction (in Molecular Biology); Introduction to statistics and bioinformatics for the analysis of large-scale biological data; Population-based Association Analysis; Introduction to genetic epidemiology; Videregående statistikk: Analyse av repeterte/korrelerte kategoriske data; Analyse av repeterte / korrelerte målinger; New statistical methods for causal inference; Videregående kurs i epidemiologiske metoder.
- **PhD-courses for students at UiO only.** Introductory course to the medical PhD program, INTRO II; Introductory course in statistics; Logistic regression, survival analysis and Cox-regression.
- **Courses for the professional study programme for medicine.**
- **Courses in the Master's Programme in Clinical Nutrition.**
- **Courses in the Master's Programme in International Community Health.**

OCBE in a nutshell: **advising activity**

What is our advising activity?

OCBE permanent staff have a duty (and OCBE PhDs and postdocs are happily involved) in helping, advising or supervising the research needs in biostatistics, epidemiology and health economics of all employees at the Faculty of Medicine (UiO) and at the OUS (or at other hospitals in HSØ).

OCBE in a nutshell: **advising activity**

What is our advising activity?

OCBE permanent staff have a duty (and OCBE PhDs and postdocs are happily involved) in helping, advising or supervising the research needs in biostatistics, epidemiology and health economics of all employees at the Faculty of Medicine (UiO) and at the OUS (or at other hospitals in HSØ).

This help is mostly important for PhD students in medicine, and for clinical grant applications that require a statistical analysis part.
Approx 12k hours of advising per year.

OCBE in a nutshell: **advising activity**

What is our advising activity?

OCBE permanent staff have a duty (and OCBE PhDs and postdocs are happily involved) in helping, advising or supervising the research needs in biostatistics, epidemiology and health economics of all employees at the Faculty of Medicine (UiO) and at the OUS (or at other hospitals in HSØ).

This help is mostly important for PhD students in medicine, and for clinical grant applications that require a statistical analysis part.
Approx 12k hours of advising per year.

Types of advising (in increasing order of OCBE involvement)

1. OCBE policlinic support;
2. single project support;
3. joint collaborative, long-term research projects.

OCBE in a nutshell: **advising activity**

What is our advising activity?

OCBE permanent staff have a duty (and OCBE PhDs and postdocs are happily involved) in helping, advising or supervising the research needs in biostatistics, epidemiology and health economics of all employees at the Faculty of Medicine (UiO) and at the OUS (or at other hospitals in HSØ).

This help is mostly important for PhD students in medicine, and for clinical grant applications that require a statistical analysis part.
Approx 12k hours of advising per year.

Types of advising (in increasing order of OCBE involvement)

1. OCBE policlinic support;
2. single project support;
3. joint collaborative, long-term research projects.

`https://www.med.uio.no/imb/english/research/centres/ocbe/advising/`

AI @ OCBE – innovative methods

Most people at OCBE spend their day doing **cutting edge research in AI:** statistical learning, machine learning, intensive computational methods

AI @ OCBE – innovative methods

Most people at OCBE spend their day doing **cutting edge research in AI:** statistical learning, machine learning, intensive computational methods
These are applied in all areas of biostatistics

AI @ OCBE – innovative methods

Most people at OCBE spend their day doing **cutting edge research in AI:** statistical learning, machine learning, intensive computational methods
These are applied in all areas of biostatistics

Model-based approach

Model based methods offer a valid framework for inference: very critical in biological / clinical applications!

AI @ OCBE – innovative methods

Most people at OCBE spend their day doing **cutting edge research in AI:** statistical learning, machine learning, intensive computational methods
These are applied in all areas of biostatistics

Model-based approach

Model based methods offer a valid framework for inference: very critical in biological / clinical applications!

In particular, Bayesian methods embed prior information & available data in a unified framework: neat approach to include strong biological knowledge into the model, which is then modified / updated according to the data.

AI @ OCBE – computations

- Everybody at OCBE develops his own software: our innovative AI methods require us to code substantially from scratch, and eventually they become "black boxes" methods which the community can use.

AI @ OCBE – computations

- Everybody at OCBE develops his own software: our innovative AI methods require us to code substantially from scratch, and eventually they become "black boxes" methods which the community can use.
- Model-based inference is expensive, both for classical statisticians (need for cross-validation, bootstrapping, ...), and for Bayesians (MCMC is not easily parallelizable).

AI @ OCBE – computations

- Everybody at OCBE develops his own software: our innovative AI methods require us to code substantially from scratch, and eventually they become "black boxes" methods which the community can use.
- Model-based inference is expensive, both for classical statisticians (need for cross-validation, bootstrapping, ...), and for Bayesians (MCMC is not easily parallelizable).
- Take home message: there is no free lunch! Models are expensive, BUT valid inference is worth the effort...

AI @ OCBE – computations

- Everybody at OCBE develops his own software: our innovative AI methods require us to code substantially from scratch, and eventually they become "black boxes" methods which the community can use.
- Model-based inference is expensive, both for classical statisticians (need for cross-validation, bootstrapping, ...), and for Bayesians (MCMC is not easily parallelizable).
- Take home message: there is no free lunch! Models are expensive, BUT valid inference is worth the effort...

What about the advising activity?

We use traditional statistical methods and/or "standard" black box AI techniques

AI @ OCBE – computations

- Everybody at OCBE develops his own software: our innovative AI methods require us to code substantially from scratch, and eventually they become "black boxes" methods which the community can use.
- Model-based inference is expensive, both for classical statisticians (need for cross-validation, bootstrapping, ...), and for Bayesians (MCMC is not easily parallelizable).
- Take home message: there is no free lunch! Models are expensive, BUT valid inference is worth the effort...

What about the advising activity?

We use traditional statistical methods and/or "standard" black box AI techniques BUT we often still need a server (TSD) because of data issues (see the next slide).

AI @ OCBE – IT use

Typical requirements on IT (hardware and software) for many of our projects:

- memory for huge datasets: clinical registries, genomic, study cohorts;
- computational power for fitting the model;
- sensitive data requirements (privacy issues with patients);
- (specifically for advising) access to data in funded projects we are not responsible for.

AI @ OCBE – IT use

Typical requirements on IT (hardware and software) for many of our projects:

- memory for huge datasets: clinical registries, genomic, study cohorts;
- computational power for fitting the model;
- sensitive data requirements (privacy issues with patients);
- (specifically for advising) access to data in funded projects we are not responsible for.

Some statistics from a recent survey...

- **software use:**
  80% R, 20% Matlab, 25% C++, 28% Phyton, SPSS and STATA
- **open-source packages production:** 20% often, 30% rarely or never, the rest does not apply
- **server use:** 64% run analyses on a server (increasing!) specifically, of those using a server: 80% use abel, 67% use TSD, 27% use med-biostat
- **sensitive data:** 52% of OCBE staff (increasing!) encounters this issue

AI @ OCBE – servers

- **"standard" TSD:** standard TSD facility for storing sensitive data; people at OCBE mostly use it for advising (when they only perform basic statistical analyses but still have sensitive data).

AI @ OCBE – servers

- **"standard" TSD:** standard TSD facility for storing sensitive data; people at OCBE mostly use it for advising (when they only perform basic statistical analyses but still have sensitive data).
- **ibm-frigessi on TSD:*** bought by the Department/BigInsight to perform HPC on TSD. Used for research, since we need computing power for innovative methods to be used on sensitive data. Linux (Red Hat Enterprise Linux 6.10) machine with 144 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.

AI @ OCBE – servers

- **"standard" TSD:** standard TSD facility for storing sensitive data; people at OCBE mostly use it for advising (when they only perform basic statistical analyses but still have sensitive data).
- **ibm-frigessi on TSD:*** bought by the Department/BigInsight to perform HPC on TSD. Used for research, since we need computing power for innovative methods to be used on sensitive data. Linux (Red Hat Enterprise Linux 6.10) machine with 144 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.
- **med-biostat:** large-memory computing server for people at OCBE. Used for research, when no sensitive data. Linux (Red Hat Enterprise Linux 7) machine with 40 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.

AI @ OCBE – servers

- **"standard" TSD:** standard TSD facility for storing sensitive data; people at OCBE mostly use it for advising (when they only perform basic statistical analyses but still have sensitive data).
- **ibm-frigessi on TSD:*** bought by the Department/BigInsight to perform HPC on TSD. Used for research, since we need computing power for innovative methods to be used on sensitive data. Linux (Red Hat Enterprise Linux 6.10) machine with 144 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.
- **med-biostat:** large-memory computing server for people at OCBE. Used for research, when no sensitive data. Linux (Red Hat Enterprise Linux 7) machine with 40 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.
- **abel, freebio:** USIT knows everything about these :)

AI @ OCBE – servers

- **"standard" TSD:** standard TSD facility for storing sensitive data; people at OCBE mostly use it for advising (when they only perform basic statistical analyses but still have sensitive data).
- **ibm-frigessi on TSD:*** bought by the Department/BigInsight to perform HPC on TSD. Used for research, since we need computing power for innovative methods to be used on sensitive data. Linux (Red Hat Enterprise Linux 6.10) machine with 144 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.
- **med-biostat:** large-memory computing server for people at OCBE. Used for research, when no sensitive data. Linux (Red Hat Enterprise Linux 7) machine with 40 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.
- **abel, freebio:** USIT knows everything about these :)
- **servers used by the PIL group:** more during the next talk!

AI @ OCBE – servers

- **"standard" TSD:** standard TSD facility for storing sensitive data; people at OCBE mostly use it for advising (when they only perform basic statistical analyses but still have sensitive data).
- **ibm-frigessi on TSD:*** bought by the Department/BigInsight to perform HPC on TSD. Used for research, since we need computing power for innovative methods to be used on sensitive data. Linux (Red Hat Enterprise Linux 6.10) machine with 144 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.
- **med-biostat:** large-memory computing server for people at OCBE. Used for research, when no sensitive data. Linux (Red Hat Enterprise Linux 7) machine with 40 cores, and 1 TB main memory. Very similar to the Abel computing environment at UiO.
- **abel, freebio:** USIT knows everything about these :)
- **servers used by the PIL group:** more during the next talk!

*One major issue with using the ibm-frigessi is that it is in the TSD secure environment, which means that the Internet is not reachable from the server. This has implications for software availability (R packages, for instance).

AI @ OCBE – needs/specifications

- often we need huge memory, but not necessarily cores;

AI @ OCBE – needs/specifications

- often we need huge memory, but not necessarily cores;
- the local IT department at the institute is not really targeted to our needs;

AI @ OCBE – needs/specifications

- often we need huge memory, but not necessarily cores;
- the local IT department at the institute is not really targeted to our needs;
- (somehow related to both points above) we are growing rapidly, both in terms of people and of projects we are involved in: we will need more and more safe and reliable IT resources;

AI @ OCBE – needs/specifications

- often we need huge memory, but not necessarily cores;
- the local IT department at the institute is not really targeted to our needs;
- (somehow related to both points above) we are growing rapidly, both in terms of people and of projects we are involved in: we will need more and more safe and reliable IT resources;
- we need computer science expertise for specific problems (e.g. running MCMC in parallel), or in general it might be useful to have an IT support for making our codes better;

AI @ OCBE – needs/specifications

- often we need huge memory, but not necessarily cores;
- the local IT department at the institute is not really targeted to our needs;
- (somehow related to both points above) we are growing rapidly, both in terms of people and of projects we are involved in: we will need more and more safe and reliable IT resources;
- we need computer science expertise for specific problems (e.g. running MCMC in parallel), or in general it might be useful to have an IT support for making our codes better;
- we (typically) don't need licensed software.

Many thanks for your attention!

Questions?