APPLICATIONS OF MACHINE LEARNING IN HEALTHCARE

Danielle Belgrave
Imperial College London
Latent Variable Modelling

Longitudinal Data Analysis

Missing Data

Bayesian Data Analysis

Causality

Survival Models
GENERAL STRUCTURE OF THIS TUTORIAL

Some **Ground Rules**: Laying the Basis

Motivation and Framework: **Endotype Discovery**

Focus: **Learning by Example**

Basic principles of **Causality**

Tips for **Team Science**
ELEMENTS OF THE PROJECT CYCLE

Understand the problem

Understand the data

Prepare the data

Evaluate Algorithms - Cross Validation

Finalise Models
Deep Learning gives excellent results on web-scale and image datasets

DL is very data hungry

Health data collection is (generally) expensive

Difficult to represent uncertainty

Interpretability

Model-Based approaches: Focus on hypothesis generating
Randomised Control Trial: Traditional Approach to Evaluating Treatment

Population is split into 2 groups by random allocation

Outcomes for both groups are measured

Patient Group

Intervention

Control

= Cured

= Still Diseased
NEED FOR PERSONALIZED TREATMENT AND MANAGEMENT STRATEGIES

Drug NOT toxic and beneficial

Drug toxic but beneficial

Drug toxic but NOT beneficial

Drug NOT toxic and NOT beneficial

Patient Group

Same diagnosis same prescription
Legacy of non-replicated genetic epidemiology, typical of most common chronic disorders

- ★ Linkage in 1 study only
- ✭ Linkage in >1 study
To identify **subgroups** of complex disease risk or treatment outcome explained by a **distinctive underlying mechanism** ("endotypes")

Foundation of **Stratified Medicine** - seeking better-targeted interventions
“We adore chaos because we love to produce order”

M.C. ESCHER
ORDER AND CHAOS, 1950
GENERALIZED FRAMEWORK FOR IDENTIFYING DISEASE ENDOTYPES

Parsimonious description of the data inferred from what is observed.
The probabilistic model expresses general knowledge about a situation.

The inference algorithm uses the model to answer queries given evidence.

The answers to queries are framed as probabilities of different outcomes.

The evidence contains specific information about a situation.

The queries express the things that will help you make a decision.

The basic components of a probabilistic reasoning system:

- Probabilistic model
- Inference Algorithm
- Evidence
- Queries
- Answer

HETEROGENEITY IN ASTHMA

Phenotypes: Observable Manifestations of Disease

Subtypes: Different Diseases With Different Causes
To define asthma subgroups (endotypes) in a population-based birth cohort study based on both parental reports and primary care consultation of wheeze within the first 8 years of life

To identify distinct genetic and physiological markers which are associated with these phenotypes
Pr(\(y_{ij} = 1\mid x_{ij}, c_i = k\)) = \(\beta_0 + \beta_1 x_{ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \xi_k + \lambda_k \text{age} + \phi_k \text{age}^2\)

\(x_{ij}\) = age; \(x_{2ij}\) rater at time \(j\); \(x_{3ij}\) is gender Pr(\(c_i = k\)) is multinomial over \(k\) classes and independent across children
ASTHMA: A HETEROGENEOUS PHENOMENON

5 distinct latent classes with different genetic and environmental characteristics

ASTHMA SUBTYPE-DEPENDENT RESPONSE TO TREATMENT

[Graphs showing the percentage of children who received inhaled steroid per class and the percentage who had exacerbations per class by age for different asthma subtypes: No Wheeze, Transient Early Wheeze, Late-onset Wheeze, Persistent Controlled Wheeze, and Persistent Troublesome Wheeze.]

- No Wheeze (n=495)
- Transient Early Wheeze (n=121)
- Late-onset Wheeze (n=157)
- Persistent Controlled Wheeze (n=116)
- Persistent Troublesome Wheeze (n=27)
DISTINCT GENETIC PROFILE OF WHEEZE SUBTYPES

Endotype discovery may have major implications for:

- Refining disease diagnosis
- **Identifying biomarkers** that allow us to understand underlying disease mechanisms
- More **personalised treatment** and management strategies of disease
Progression of allergy:
Eczema -> Asthma -> Rhinitis

Symptoms Causally Linked

Prevention strategy:
Target children with eczema to reduce progression to asthma and rhinitis
OBJECTIVE

To capture disease heterogeneity and encapsulate different patterns of symptom progression during childhood using a probabilistic modelling approach.
THE DATA DOMAIN

Manchester Longitudinal birth cohort ~2000 children

Bristol Longitudinal birth cohort ~10000 children

Prevalence (%)

Age (Years)

Eczema
Wheeze
Rhinitis
HIDDEN MARKOV MODEL 1: INDEPENDENT PROFILES

Manchester Asthma and Allergy Study
1184 subjects

Avon Longitudinal Study of Parents and Children
8665 subjects
HIDDEN MARKOV MODEL 2: “ALLERGIC MARCH”
MODEL 3: LONGITUDINAL LATENT DISEASE PROFILE

Latent State Age 1
- Eczema Age 1
- Wheeze Age 1
- Rhinitis Age 1

Latent State Age 3
- Eczema Age 3
- Wheeze Age 3
- Rhinitis Age 3

Latent State Age 5
- Eczema Age 5
- Wheeze Age 5
- Rhinitis Age 5

Latent State Age 8
- Eczema Age 8
- Wheeze Age 8
- Rhinitis Age 8

Latent State Age 11
- Eczema Age 11
- Wheeze Age 11
- Rhinitis Age 11

Class = 1,...,k

Children (n=9801)
INFER.NET INference Architecture

Probabilistic program

Infer.NET compiler → C# compiler → Algorithm execution

Infer.NET Inference Engine

Observed values (data, priors)

Probability distributions
## SENSITIVITY TO PRIORS

<table>
<thead>
<tr>
<th>Prior on the number of pseudo-counts</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/n</td>
<td>-50177</td>
<td>-49030</td>
<td>-48297</td>
<td>-47774</td>
<td>-47367</td>
<td>-47130</td>
<td>-46989</td>
<td>-47109*</td>
</tr>
<tr>
<td>2/n</td>
<td>-50200</td>
<td>-49104</td>
<td>-48310</td>
<td>-47797</td>
<td>-47357</td>
<td>-47143</td>
<td>-46994</td>
<td>-47334*</td>
</tr>
<tr>
<td>1</td>
<td>-49920</td>
<td>-48448</td>
<td>-47506</td>
<td>-46930</td>
<td>-46845</td>
<td>-46658</td>
<td>-46503</td>
<td>-46424*</td>
</tr>
<tr>
<td>2</td>
<td>-49920</td>
<td>-48448</td>
<td>-47506</td>
<td>-46930</td>
<td>-46845</td>
<td>-46733</td>
<td>-46596</td>
<td>-46431*</td>
</tr>
</tbody>
</table>
# Posterior Probability of Class Membership

<table>
<thead>
<tr>
<th>Class</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.924</td>
<td>0.783</td>
<td>0.805</td>
<td>0.805</td>
<td>0.756</td>
<td>0.805</td>
<td>0.846</td>
</tr>
<tr>
<td>0.943</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The **Allergic March** reflects patterns at the population level, rather than the natural covariance of symptoms within individuals’ life courses.

Developmental profiles of Eczema → Asthma → Rhinitis are heterogeneous.

Only a small proportion of children follow a trajectory profile similar to that of the atopic march.
Stop the killing of beneficial bacteria

Concerns about antibiotics focus on bacterial resistance — but permanent changes to our protective flora could have more serious consequences, says Martin Blaser.

- Average child in developed countries takes 10-20 courses of antibiotics before age 18 yr

Factors affecting Early Respiratory Colonisation

- Antibiotics
- Delivery Method
- Bowel Colonisation
- Nasopharyngeal Colonisation
- Feeding Method

Impact of Early Respiratory Colonisation

- Risk of Bronchiolitis
- Long Term Risk of Asthma
- Risk of Chronic Lung Disease in Preterm Infants

Neonatal Airway and Lung Bacterial Colonisation
OTU's are used to categorize bacteria based on sequence similarity.

### OTU ID

<table>
<thead>
<tr>
<th>#OTU ID</th>
<th>D53<del>DRun8</del>24moS wab</th>
<th>D283<del>DRun15</del>24moS wabs</th>
<th>DI73<del>DRun15</del>24moSw abs</th>
<th>D131<del>DRun15</del>24moSw abs</th>
<th>D225<del>DRun15</del>24moS wabs</th>
<th>D98<del>DRun15</del>24moSw abs</th>
</tr>
</thead>
<tbody>
<tr>
<td>New.ReferenceOTU75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New.ReferenceOTU76</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>New.ReferenceOTU77</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New.ReferenceOTU8</td>
<td>64</td>
<td>14</td>
<td>20</td>
<td>23</td>
<td>57</td>
<td>96</td>
</tr>
<tr>
<td>New.ReferenceOTU9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New.ReferenceOTU10</td>
<td>0</td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New.ReferenceOTU11</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New.ReferenceOTU12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New.ReferenceOTU13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New.ReferenceOTU14</td>
<td>60</td>
<td>78</td>
<td>29</td>
<td>8</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

### Taxonomic Composition

<table>
<thead>
<tr>
<th>superkingdom</th>
<th>phylum</th>
<th>class</th>
<th>order</th>
<th>family</th>
<th>genus</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroidia</td>
<td>Bacteroidales</td>
<td>Prevotellaceae</td>
<td>Prevotella</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Oribacterium</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Actinomycetales</td>
<td>Actinomycetaceae</td>
<td>Actinomyces</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Actinobacteria</td>
<td>Bacteroidia</td>
<td>Bacteroidales</td>
<td>Porphyromonadaceae</td>
<td>Porphyromonas</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Actinobacteria</td>
<td>Fusobacteria</td>
<td>Fusobacteriales</td>
<td>Leptotrichiaceae</td>
<td>uncultured</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>Pasteurellales</td>
<td>Pasteurellaceae</td>
<td>Haemophilus</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Negativicutes</td>
<td>Selenomonadales</td>
<td>Veillonellaceae</td>
<td>Veillonella</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>Pasteurellales</td>
<td>Pasteurellaceae</td>
<td>Actinobacillus</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroidia</td>
<td>Bacteroidales</td>
<td>Porphyromonadaceae</td>
<td>Porphyromonas</td>
<td>uncultured bacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Betaproteobacteria</td>
<td>Neisseriales</td>
<td>Neisseriaceae</td>
<td>Neisseria</td>
<td>uncultured bacterium</td>
</tr>
</tbody>
</table>
CORRELATION NETWORK ANALYSIS
EVOLUTION OF MICROBIOME PROFILE OVER TIME
**MICROBIOME PROFILE AND RESPIRATORY DISEASE**

Bacterial Composition of 1,021 Nasopharyngeal Aspirates Collected from 234 Infants during Periods of Respiratory Health and Disease

Clustering based on the 6 most common genera

---

AN AGE-OLD PROBLEM...

12.7 million people discover they have cancer each year

7.6 million people die from cancer each year

30 - 40% of these deaths can be prevented
Lack of tools for early detection and diagnosis

Cancer cells, even within the same tumor, are heterogeneous—that is, differences exist between the individual cells.
Aim: To determine the difference between cancerous gene expression in tumour cells vs normal, non-cancerous tissues to obtain better insight into the disease pathology.

To create a generalizable framework for new cancer types without the redesign of new features.
Delayed ICU admission is correlated with mortality

Ignoring correlations among vital signs, history and patient heterogeneity

Risk scoring methodology can confer huge clinical and social benefits on a massive number of critically ill inpatients who exhibit adverse outcomes including, but not limited to, cardiac arrests, respiratory arrests, and septic shocks.
A MULTI-TASK GAUSSIAN PROCESS MODEL FOR ICU ADMISSION

Results reflect the importance of adopting the concepts of personalized medicine in critical care settings; significant accuracy and timeliness gains can be achieved by accounting for the patients' heterogeneity.

Personalisation: Identify Endotypes via Latent Class Model

<table>
<thead>
<tr>
<th>Plausibility</th>
<th>Consistency</th>
<th>Temporality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strength</td>
<td>Specificity</td>
<td>Change in</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Risk Factor</td>
</tr>
</tbody>
</table>
A variable that changes the impact of one variable on another
A **mechanism** by which one variable affects another variable

**Predictor** (Independent Variable) → **Mediator** → **Outcome** (Dependent Variable)
**TESTING MEDIATION**

**Step 1:** Independent Variable $\rightarrow$ Dependent Variable

**Step 2:** Independent Variable $\rightarrow$ Mediator

**Step 3:** Mediator $\rightarrow$ Dependent Variable

**Step 4:** Effect of Independent Variable on Dependent Variable is significantly reduced by controlling for the mediator:


INSTRUMENTAL VARIABLE (IV) ESTIMATION

Allows for consistent, unbiased estimation when the explanatory variables (covariates) are correlated with the error term in a regression model

Used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment.
Scenarios:

Change in the dependent variable change the value of at least one of the covariates (reverse causation)

Omitted variables that affect both the dependent and independent variables

Covariates are subject to measurement error
MEDIATION WITH INSTRUMENTAL VARIABLES

Instruments -> Error
Random Allocation -> Error
Covariates -> Error
Error -> U
Error -> Outcomes
An instrumental variable is:

1. Strongly predictive of the mediating variable
2. Has no direct effect on the outcome except through the mediator
3. Does not share common causes with the outcome

Randomisation, where available, often satisfies this criteria when accounting for departures from randomised treatment.

“Correlation and Causality” by David Kenny (1979)
Efficacy and mechanism evaluation: Causal framework for investigating who medications work for.

Using the treatment by marker interaction as an instrument.
EFFICACY AND MECHANISM EVALUATION: CANCER

- Genetic Marker
- Treatment
- Prognostic biomarker (risk factor)
- Tumor Size
- Outcome (Survival)
“There is less attention paid to the more immediate problem of how we prevent these programs from amplifying the inequalities of our past and affecting the most vulnerable members of our society.”

ML IN HEALTH: THERE IS STILL A LOT THAT NEEDS TO BE DONE...
The key to collaboration is effective communication

REFLECTIONS ON TEAM SCIENCE

Belgrave et al. Disaggregating asthma: Big investigation versus big data. Journal of Allergy and Clinical Immunology 139.2 (2017): 400-407.
Think deeply about the clinical context. Find solutions which are specific to the problem.

Good science is about merging different schools of thought for developing the bigger picture.

Data driven approach + Domain Knowledge = Holistic Approach to science

**REFERENCES ON TEAM SCIENCE**

Principled epidemiology + Biostatistics + Machine Learning = Heuristic Blend of Tools for understanding causality and clinical relevance

REFLECTIONS ON TEAM SCIENCE

1. **Team Science**: Discoveries about healthcare, not hypothesised a priori, have been made by experts explaining structure learned from data by algorithms tuned by those experts.

2. Heuristic blend of **biostatistics** and **machine-learning** reveals more than either method individually.

3. An ML approach to extracting knowledge from information in healthcare requires persistent integration of
   Data
   Methods
   Expertise
THE ROAD AHEAD...
Assumptions:
Children in the same class have similar transitions of symptoms over time

```csharp
public ClusterSimpleChain(int numYears)
{
    ...
    probState0 = Variable.Array<double>(k).Named("probState0");
    probState0Prior = Variable.Array<Beta>(k).Named("probState0Prior");
    probState0[k] = Variable<double>.Random(probState0Prior[k]);

    for (int y = 0; y < numYears; y++)
    {
        #if clusterQ
            Q_T[y] = Variable.Array(Variable.Array<double>{s}, k).Named("Q_T" + y);
            Q_F[y] = Variable.Array(Variable.Array<double>{s}, k).Named("Q_F" + y);
            QTPriorArr[y] = Variable.Array(Variable.Array<Beta>{s}, k).Named("QTPriorArr" + y);
            QFPriorArr[y] = Variable.Array(Variable.Array<Beta>{s}, k).Named("QFPriorArr" + y);
            Q_T[y][k][s] = Variable<double>.Random(QTPriorArr[y][k][s]);
            Q_F[y][k][s] = Variable<double>.Random(QFPriorArr[y][k][s]);
        #else
            Q_T[y] = Variable.Array<double>{s}.Named("Q_T" + y);
            Q_F[y] = Variable.Array<double>{s}.Named("Q_F" + y);
            QTPriorArr[y] = Variable.Array<Beta>{s}.Named("QTPriorArr" + y);
            QFPriorArr[y] = Variable.Array<Beta>{s}.Named("QFPriorArr" + y);
            Q_T[y][s] = Variable<double>.Random(QTPriorArr[y][s]);
            Q_F[y][s] = Variable<double>.Random(QFPriorArr[y][s]);
        #endif
        ...
    }
```