Applied

* Roche
* Merck
* AstraZeneca
* Novartis – jobs in india – didn’t apply yet

Should apply

* Bayer
* Sanofi
* Amgen
* Baxter – jobs in india
* BD
* IBM
* Google
* Microsoft
* MSKCC
* Broad

Work visa issue currently

* Pfizer – jobs mostly in the US
* Biogen – mostly US jobs
* Eli Lilly – nothing good outside of the US – may 22 2018
* Bristol-Myers Squibb – mostly US – some in China
* Johnson and Johnson
* GSK – need work permit
* Gilead – mostly us jobs - Only candidates in possession of a valid EU work permit will be considered.
* Abott – all about diabetes
* Abbvie – abroad no stats jobs – except in Germany – you need excellent german

AstraZeneca Pfizer GSK Roche Sanofi Amgen Novartis Merck

US also – big companies like roche

1. The cv and cover letter wont get any better – as soon as you’re within a few days of being interview ready – apply
2. Cover letter – tailor it by saying – your pipeline has so many drugs – or cancer drugs particularly in immune oncology
3. Info interviews – known people and extended -pfizer– unknown biostats, comp bio people – unknown duke alumni – see where they are – two things stuck out for me – the guy had a good background – specific achievements of mine from linked listed, I was proud of these – and I was very happy he contacted me for advice
4. Do info interviews – try duke and comp bio/stat – if you cant find – duke and oncology – ask them to recommend someone I can speak to
5. Now before leaving for merck is the best time to talk to people who go that award if they are working in Europe – apply for those jobs first

Reach out to people to make a connection

1. If nothing works out – after paper draft is done – apply for post docs/research scientist
2. In that post doc keep networking and apply again – you have a good cv and great skills, you’ll get through

Tech leadership cover letter:

Strong time management skills; able to effectively organise and manage a variety of tasks across different projects.

-through effective communication and influence.

• Excellent interpersonal and communication skill, including:  
o Demonstrated ability building and maintaining strong working relationships

Self-motivated and independent worker.

Linkedin reco:

Have novel innovative ideas - phd

You can identify whats important even when it’s a new field

You’ll teach yourself that – self motivated – and very quickly -Manage to finish the job within time – phd

result

Leadership

* Identifying whats important big picture – details
* Adapt to changes
* Motivating
* Resolving conflicts quickly
* Make sure results are delivered on time

1. learnt a lot of new techniques – knows DNA sequence, RNA seq- identifying significant genes and pathways
2. even if it is a new area – or a new problem – she can quickly find the best solution to the problem
3. Led a project on pan cancer – led people
4. hard working – diligent
5. can work independently – figures out solutions to problems – thinks about how creatively and efficiently

 (I) Statistical and computational: R, SAS, JMP, Spotfire, and Matlab; (II) Progamming environment: Unix/Linux; (III) Programming languages: Python, Perl; (IV) Bioinformatics: Bioconductor, IPA; (V) Database management systems.

* AstraZeneca

<https://job-search.astrazeneca.com/job/gothenburg/associate-principal-scientist-biostatistics-bioinformatics/7684/6197770> - did it expire?

<https://job-search.astrazeneca.com/job/cambridge/genome-analyst/7684/7036343> - applied

<https://job-search.astrazeneca.com/job/waltham/oncology-bioinformatics-data-science-roles/7684/6721979> - expired?

<https://job-search.astrazeneca.com/job/cambridge/bioinformatics-pipeline-manager/7684/6259022> - expired?

<https://job-search.astrazeneca.com/job/waltham/post-doc-fellow-optimizing-clinical-development-programs-using-bayesian-statistic-and-machine-learn/7684/7456554?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-023768&ss=paid>

<https://job-search.astrazeneca.com/job/cambridge/principal-health-informatics-scientist/7684/6545242?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-019441&ss=paid> - not such a great fit

<https://job-search.astrazeneca.com/job/cambridge/senior-data-scientist-12-months-fixed-term-contract/7684/6321786?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-017649&ss=paid> - applied

<https://job-search.astrazeneca.com/job/waltham/principal-scientist-ngs-informatics-oncology/7684/7469612?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-024121&ss=paid> - expired?

<https://job-search.astrazeneca.com/job/gothenburg/associate-principal-scientist-biostatistics-bioinformatics/7684/6197770?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-016500&ss=paid> - expired?

<https://job-search.astrazeneca.com/job/gothenburg/associate-director-quantitative-biology/7684/7493975?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-024292&ss=paid> - expired?

<https://job-search.astrazeneca.com/job/cambridge/senior-statistician-oncology/7684/5243249?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-011571&ss=paid> - applied

<https://job-search.astrazeneca.com/job/cambridge/principal-statistician-statistical-science-director/7684/5218707?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-010997&ss=paid> - applied

<https://job-search.astrazeneca.com/job/vastra-gotaland/data-scientist/7684/7505605?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-024375&ss=paid> - applied

<https://job-search.astrazeneca.com/job/cambridge/senior-principal-data-operations-programmer-early-clinical-development/7684/7520584?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-024504&ss=paid> - very sas heavy

<https://job-search.astrazeneca.com/job/cambridge/postdoc-fellow-characterizing-immune-tumour-microenvironment/7684/7417161?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-023563&ss=paid>

https://job-search.astrazeneca.com/job/gothenburg/post-doc-fellow-regulation-of-rna-metabolism-in-lung-epithelium-by-hur/7684/7479619?utm\_source=jobalert&utm\_medium=email&utm\_campaign=GlobalTB\_jobalert&utm\_content=R-024184&ss=paid

<https://job-search.astrazeneca.com/job/cambridge/post-doc-fellow-identifying-oncogenic-regulators-through-networks/7684/7417162?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-023490&ss=paid>

<https://job-search.astrazeneca.com/job/cambridge/post-doc-fellow-prostate-cancer-genomics/7684/7417147?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-023474&ss=paid>

<https://job-search.astrazeneca.com/job/waltham/post-doc-fellow-optimizing-clinical-development-programs-using-bayesian-statistic-and-machine-learn/7684/7456554?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-023768&ss=paid>

https://job-search.astrazeneca.com/category/scientific-jobs/7684/40091/1

R-011571 Senior Statistician - Oncology

R-026637 Machine Learning/Artificial Intelligence Data Scientist

R-024461 Senior/Associate Principal Statistician

R-024375 Data Scientist

R-021796 Genome Analyst

R-023124 Senior Scientist, Bioinformatics, Oncology

R-010997 Principal Statistician / Statistical Science Director

R-017649 Senior Data Scientist - 12 months fixed term contract

 R-025307 Associate Principal Scientist

 R-016500 Combined Biostatistics and Bioinformatics Expert – cover letter gaffe!!!!!!!! – I withdrew the application

R-027565 Senior Statistician, Quantitative Biology

R-020222 Informatics Scientist

R-027314 Data & Analysis Engineer - Genomics

R-027247 Data Scientist - Global Medical Affairs

R-020639 Information Management Specialist, Oncology

Info interviews:

Margie Li

2nd degree connection2nd

Oncology Strategy Director at AstraZeneca

# Mercedes Vazquez Chantada

3rd degree connection3rd

## Senior Research Scientist at AstraZeneca

### **AstraZeneca**

### **Universidad del País Vasco/Euskal Herriko Unibertsitatea**

### **Cambridge, United Kingdom**

# Sriram Nagaraj

3rd degree connection3rd

## Senior Director, Immuno-Oncology Marketing, US Oncology at AstraZeneca

### **AstraZeneca**

### **Duke University - The Fuqua School of Business**

### **Washington D.C. Metro Area**

# Judy Schreiber, PhD, RN

3rd degree connection3rd

## Pan-Tumor Immuno-Oncology Medical Science Liaison

### **AstraZeneca**

### **University of Kentucky**

# Zapporah Young, PhD

3rd degree connection3rd

## Postdoctoral Research Scientist at AstraZeneca

### **AstraZeneca**

### **University of Michigan**

# Nicholas D'Amato

3rd degree connection3rd

## Oncology MSL (Women's Cancer and I/O)

### **AstraZeneca**

### **Duke University**

Merck

<https://www.emdgroup.com/en/careers/jobs/169158>

<https://www.emdgroup.com/en/careers/jobs/175471>

<https://www.emdgroup.com/en/careers/jobs/176369>

<https://www.emdgroup.com/en/careers/jobs/175770>

<https://www.emdgroup.com/en/careers/jobs/175158>

<https://www.emdgroup.com/en/careers/jobs/174974>

<https://www.emdgroup.com/en/careers/jobs/172259>

<https://www.emdgroup.com/en/careers/jobs/171079> - post doc 12 months after phd – and this is experimental

maybe:

https://www.emdgroup.com/en/careers/jobs/175552

<https://www.emdgroup.com/en/careers/jobs/175539>

https://www.emdgroup.com/en/careers/jobs/175156

# AVISEK DEYATI, PhD

2nd degree connection2nd

## Principal Investigator at Biocon Bristol Myers Squibb R&D Center

# Bijan Zakeri

2nd degree connection2nd

## Senior Scientist in Biologics Discovery at EMD Serono

### **Greater Boston Area**

Pfizer

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/USA---CA---La-Jolla/Sr-Scientist-Computational-Biology--Cancer-Biology_2934709-2?utm_source=Indeed&utm_medium=organic&utm_campaign=Indeed>

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/United-States---Connecticut---Groton/Computational-Toxicologist-Biologist--Sr-Scientist_4689616-1>

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/United-States---Massachusetts---Cambridge/Postdoctoral-Fellow--Systems-Modeling_4692921>

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/United-States---Connecticut---Groton/Computational-Toxicologist-Biologist--Sr-Scientist_4689616-1>

4692921

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/United-States---Massachusetts---Cambridge/Postdoctoral-Fellow--Systems-Modeling_4692921>

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/USA---CA---Rinat/Computational-Biologist--Immuno-oncology_1610649-1>

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/United-States---Connecticut---Groton/Computational-Toxicologist-Biologist--Sr-Scientist_4689616-1>

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/USA---CA---La-Jolla/Sr-Scientist-Computational-Biology--Cancer-Biology_2934709-2>

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/United-States---New-York---New-York-City/Senior-Data-Scientist_4697713-1>

<https://pfizer.wd1.myworkdayjobs.com/en-US/PfizerCareers/job/United-States---Connecticut---Groton/Computational-Toxicologist-Biologist--Sr-Scientist_4689616-1>

Genentech:

<https://www.gene.com/careers/detail/3231309824/Sr-Scientist-Group-Leader-Bioinformatics?src=JB-12568>

Sanofi:

<https://www.indeed.com/viewjob?jk=78ee5cd6d2833867&q=computational+biology+(scientist+or+computational+or+biology+or+bioinformatics)+$115,000%2B&tk=1caebtmp236pv8uq&from=ja&alid=5ac689532e7c4721727a76f0&utm_source=jobseeker_emails&utm_medium=email&utm_campaign=job_alerts&rgtk=1caebtmp236pv8uq>

<https://en.jobs.sanofi.com/job/cambridge/precision-immunology-senior-data-scientist/20873/7699271>

<https://en.jobs.sanofi.com/job/cambridge/senior-data-scientist/20873/7625865>

<https://en.jobs.sanofi.com/job/cambridge/rwe-data-analyst/20873/7608622>

<https://en.jobs.sanofi.com/job/cambridge/sr-data-scientist/20873/7461090>

<https://en.jobs.sanofi.com/job/cambridge/data-scientist/20873/7451494>

<https://en.jobs.sanofi.com/job/cambridge/computational-systems-immunologist-precision-immunology-cluster-immunology-and-inflammation-therape/20873/7277755>

<https://en.jobs.sanofi.com/job/cambridge/precision-immunology-data-scientist-immunology-and-inflammation-therapeutic-area/20873/7277763>

<https://en.jobs.sanofi.com/job/swiftwater/senior-biostatistician/20873/6196549>

Applied: https://sanofi.wd3.myworkdayjobs.com/en-US/SanofiCareers/userHome

R2451795 - Precision Immunology Principal Data Scientist at Sanofi

R2451271 - Senior Data Scientist at Sanofi.

R2456742 - Scientist (m/f) Artificial Intelligence at Sanofi. Y

3rd degree connection3rd

## Head of Project Management Group, Oncology Development at Sanofi Genzyme

### **Sanofi Genzyme**

### **Carnegie Mellon University - Tepper School of Business**

# Douglas Keller

3rd degree connection3rd

## Global Head for Development Projects, Preclinical Safety at Sanofi

### **Sanofi**

### **Duke University**

# Lin Liu

3rd degree connection3rd

## Principal Scientist at Sanofi

### **Sanofi**

### **Duke University**

# Scott Lee

3rd degree connection3rd

## Principal Research Investigator at sanofi-aventis

### **sanofi-aventis**

### **Duke University**

J&J

[**https://www.indeed.com/viewjob?jk=6d52d4d93247b910&q=computational+biology+(scientist+or+computational+or+biology+or+bioinformatics)+$115,000%2B&tk=1caebtmp236pv8uq&from=ja&alid=5ac689532e7c4721727a76f0&utm\_source=jobseeker\_emails&utm\_medium=email&utm\_campaign=job\_alerts&rgtk=1caebtmp236pv8uq**](https://www.indeed.com/viewjob?jk=6d52d4d93247b910&q=computational+biology+(scientist+or+computational+or+biology+or+bioinformatics)+$115,000%2B&tk=1caebtmp236pv8uq&from=ja&alid=5ac689532e7c4721727a76f0&utm_source=jobseeker_emails&utm_medium=email&utm_campaign=job_alerts&rgtk=1caebtmp236pv8uq)

**Bristol Myers Squib**

[**https://www.indeed.com/viewjob?jk=6bb1fcb499a910d1&q=computational+biology+(scientist+or+computational+or+biology+or+bioinformatics)+$115,000%2B&tk=1caebtmp236pv8uq&from=ja&alid=5ac689532e7c4721727a76f0&utm\_source=jobseeker\_emails&utm\_medium=email&utm\_campaign=job\_alerts&rgtk=1caebtmp236pv8uq**](https://www.indeed.com/viewjob?jk=6bb1fcb499a910d1&q=computational+biology+(scientist+or+computational+or+biology+or+bioinformatics)+$115,000%2B&tk=1caebtmp236pv8uq&from=ja&alid=5ac689532e7c4721727a76f0&utm_source=jobseeker_emails&utm_medium=email&utm_campaign=job_alerts&rgtk=1caebtmp236pv8uq)

[**https://www.indeed.com/viewjob?jk=ae7b3c7e44921efa&q=computational+biology+(scientist+or+computational+or+biology+or+bioinformatics)+$115,000%2B&tk=1caebtmp236pv8uq&from=ja&alid=5ac689532e7c4721727a76f0&utm\_source=jobseeker\_emails&utm\_medium=email&utm\_campaign=job\_alerts&rgtk=1caebtmp236pv8uq**](https://www.indeed.com/viewjob?jk=ae7b3c7e44921efa&q=computational+biology+(scientist+or+computational+or+biology+or+bioinformatics)+$115,000%2B&tk=1caebtmp236pv8uq&from=ja&alid=5ac689532e7c4721727a76f0&utm_source=jobseeker_emails&utm_medium=email&utm_campaign=job_alerts&rgtk=1caebtmp236pv8uq)

[**https://www.bms.com/job-seekers/job-search-results/job-details.html?requisitionId=R1505207**](https://www.bms.com/job-seekers/job-search-results/job-details.html?requisitionId=R1505207)

[**https://www.bms.com/job-seekers/job-search-results/job-details.html?requisitionId=R1504607**](https://www.bms.com/job-seekers/job-search-results/job-details.html?requisitionId=R1504607)

**GSK:**

Statistician/Principal Statistician – applied

Nonclinical\_Biostatistician

Statistician/Principal Statistician

Statistical Analyst for Non-Clinical Statistics Vaccines R&D (m/f)

Req ID: WD156348

**Roche**

* Clinical diagnostic company

Dear Ms. Smith:

Thank you for the opportunity to interview with you. I am excited to learn more about your organization and to discuss my interests and experiences. I am available during the following times for an interview.

Monday, Nov 19th 9:00am–Noon EST (6:00‐9:00am PST)

Tuesday, Nov 20th 10:00am – 2:00pm EST (7:00‐11:00am PST)

My phone number is 714‐520‐5060. I look forward to speaking with you soon.

Sincerely,

# Sarah O'Brien

2nd degree connection2nd

## Senior Associate Scientist at Amgen

### **San Francisco Bay Area**

# Michael Becker

2nd degree connection2nd

## Head of Laboratory at Bayer

### **Cologne Area, Germany**

# Frances Vu

2nd degree connection2nd

## Regional Medical Advisor at Sanofi Genzyme

### **Canada**

# Christine Ju

2nd degree connection2nd

## Senior Biostatistician at Roche Sequencing Solutions

### [Amruta Yedekar](https://www.linkedin.com/in/amrutayedekar/)

### [Coulter Knapp Coulter Knapp is a Premium member](https://www.linkedin.com/in/coulterknapp/)

Data Scientist at Roche

San Francisco Bay Area

# Shamil Sadikhov

3rd degree connection3rd

## Senior Biostatistician at Roche

* Lead novel target identification initiatives in CVMD domain anchored in human genetics by employing systems level ‘omics’ and biomarker data integration, analysis and mining.
* Senior data scientist gmd – all about picking the right patients –also includes which biomarkers are associated with which patient from clinical study outcomes - helps to know which genes/pathways affected –
* Say sequence – structure – combination of biomarkers
* Know what to say – and specific question – about post doc, phd
* Know what to say – and specific question – about mini projects on your cv
* Know what to say – and specific question – about all the methods and stat tests on ppt
* General
* Specific to the position as in job description – so like clinical trial stuff and nlp
* Coding
* Know everything on your cv
* Know what to say – why you did this and not something else
* Know answers to all the gen questions too
* Know answers to all things on the job descrption
* Be prepared for your talk

Like gavin said – talk to them like a colleague - Think you are contributing to their team – and they want you as well – impress though

Initial interview was a phone-based interview. I explained my current role at the time and how it would fit into the posted job's roles and responsibilities – resume discussion – some behavioral

Tech questions - This is what the test was – and I used it This was the problem – this is what I did – stats heavy – this was the result – adv, disadv

Statistical interview structure:

Like everyone you know some techniques awesome – and all others you’ve taken classes on – you also implemented everything!!

From taking the stats for clinical trials class on courser:

* “All” the stats for clinical trials are the standard stats so that’s great!
* Going through all the stats methods to see if theres points I didn’t know before- and I did moooost of them so that’s great!
* The problems in clinical trials itself not that important – it’s this is the data – what is the best test
* To get a sense of what problems they solve
  + – its mostly drug, placebo –or treatment1, treatment2 - this is the effect
  + – connection between smoking and lung cancer
  + patients who are fat vs not – these are the continuous distributions
* This is what I did in terms of what they need– high level this is what the problem is, approach and results – I can take you through the bioinformatics approach in more detail – this is how it is relevant to the position
* This is the method – good at – not so good at including assumptions – I actually used it – this was the problem – I used this package – these were the issues which I resolved – this was the result (star)

Then he might quiz you – this is the problem – what would you do – think for a 1min – say these are the methods – pros and cons

A few stories with high impact – told to the effect that they care

Don’t overdo it with number of stories

5/10 mins before just blank out – don’t keep revising till last minute

* Tell me about yourself
* What they do how is it relevant - that I have this background – this is how I can apply it to get them drug targets
* What ideas do they have - noncoding, epigenetic info with coding – phenotyping – existing biomarkers? – therapy with the rna structure
* Specific resume – go through everything you did
* Are they good at stats – read up
* What is your expertise
* what do you want them to know about you
* everything I did well - I have many awards
* even when new - quick learner – very intelligent
* Did they lead
* very interested in this field – not so great
  + education – i have used most stats techniques in research collaborations or classes and anything I didn’t use, I take a project and implement it on my own
  + experience - so my phd was in computer engineering and stats – I deliberately chose to apply that to oncology – that’s also why I applied for the merck innovation award

Lessons learnt from az tech interview:

* Yea its tech – but you need to be prepared for tell me about youself and gen questions as well

Lessons from first gmd interview

* Good job tell me about yourself
* Good job delivering and answering questions on ppt
* Good job saying these are areas we can best be useful and saying experimental ways to prove that
* For fit – said yea I have these skills – and that Im happy to work without getting paid on modeling on healthcare data – not sure if convincing – if someone has direct experience
* No us visa – not sure if they find a us candidate
* Good questions you asked him throughout – stopped to think if they were good questions
* Your so stuck on all the details – yea know that- but really need to focus on big picture – this is what I do – this is how I will fit in with you – yea you said biomarker – but also think about all the skills and how you can apply it to all the things they do – and why you are interested in all the things they do – you neeeeed to come across like you are interested in their projects – not the job

Lessons learnt from clinical info science principal

* Great job predicting questions!!
* Whats your ideal job

If you don’t anticipate a question – your actual opinion – or what you do is good

Take notes

My post doc was primarily a bioinformatics project

My phd was primarily biostatistics – computational and experimental validation of my models

I’m a currently a postdoc in the duke school of medicine – lead? –masters level - biostats and bioinformatics dept –

The first wgs study of bile duct cancer

For my post doc, I conceived and implemented a project that focuses on analyzing mutations in the noncoding regions of the genome so promoters and enhancers …..it was recently found that majority of disease causing mutations are in the noncoding regions - we know that even though coding reions are analyzed

– wgs sequencing stdies expensive - these regions were not analyzed so far and only now we can because we have wgs data for many diseases

I was recently involved in an international collaboration between 8 countries to analyze the first wgs study in bile duct cancer – we discovered pathways that were dysregulated in bile duct cancer– these results were also experimentally validated – and the preliminary study was published in cancer discovery. - but we didn’t find it by analyzing mutations in the coding regions alone

I subsequently expanded the analysis to 11 different cancer types – this time very robust statistical analysis and created multiple bioinformatics pipeliens - which resulted in known genes/pathways as well as novel genes/pathways correlated with these cancers.

Two orthogonal pipelines: to get genes/pathways and the other to get tfs

the bioinformatics pipeline is mostly a series of filtration steps –4 steps in particular are crucial

1. first if you take wgs data and restrict to prmoters/enhancers – most are passengers – to find driver – improve s/n - – combine with epigenetic information to see if the site is open and closed - takes high through experimental predictions that say this mutation causes this tf to be dysregulated an changesthe behavior of the gene
2. what we do to increase power of rare variant analysis is group mutations per gene or pathway – fishers exact test – 2 types of mutations – testing the effect on 2 types of cases

bernards test – slower

1. im always cognizant of the fact that you can get statistically signigicant results with very low p values but when you actually test this in the clinic, it may not work – so you want to include as much biological information as possible and do control studies - and do a series of controls (synthetic mutations and pathways) and statistical tests (fishers exact test, wilcoxon mann whiteney test) – some of them more specifically are generate synthetic mutations so mutations that are not relevant to this cancer and if the same results come up then it’s a red flag – we also generate synthetic pathways and if a tf comes up as significant

* you should make sure the gene expression is not changing because of substitutions/insertions/deletions

finally we removed as much noise as possible

1. we also do gene expression analysis both at the gene level using deseq2 and at a pathway level using gsea, gsa, ipa – this is after correcting for batch effects - to determine if that gene or pathway is significantly altered in the disease state compared to the normal state

* to improve subtype patients
* sequence is reference genome
* germline mutations

Im very excited to extend this analysis to different disease types – and to subgroups of patients -based on the genes and pathways that are affected – we can subtype patients to see which patient will likely respond to a therapy - and also help select patients for clinical trials

What Im very excited to do is take this novel anlaysis and apply it to new disease areas – im particularly excited about applying it to cv disease cause it causes the most deaths world wide – which is why I applied to this position

-other methods to do this and cvmd methods

I’m very excited about using this analysis in cvmd – this has not been done in cvmd so far - it can result in

1. find tfs dysregulated in cvrmd - transcription factors are commonly used as drug targets – in nature reviews drug discovery – article Unexplored therapeutic opportunities in the human genome – which az was involved with only 7% drug targets are tfs and that percentage can be improved

* in cv disease we know tfs that are dysregulated but this would be a comprehensive and systematic way of identifying more tfs
* AZ uses mRNA to create proteins
* - this can be used to treat even before symptoms emerge

used as drug targets- now though with wgs data and also high throughput assays that inform how tfs work - we can informaiton that was not possible before

Development of the [cardiovascular system](https://www.sciencedirect.com/topics/neuroscience/cardiovascular-system), including the heart, is a multistep process that is coordinated by a network of transcription factors

1. based on the genes and pathways that are affected – we can subtype patients to see which patient will likely respond to a therapy - and also help select patients for clinical trials -

-so the personalized medicine nature of this – will result in therapies with maximal efficacy and no or limited adverse reactions

ideally want to integrate this information with epigenetic and coding region predictions

90% of AZ pipeline focuses on precision medicine - astrazeneca with the goal of analyzing 2 million genomes by 2026 – this is the perfect place to do this type of analysis

* also az very collaborative
* you can do what you are interested in

GATK:

https://software.broadinstitute.org/gatk/documentation/article?id=11136

The panel of normals not only represents common germline variant sites, it presents commonly noisy sites in sequencing data, e.g. mapping artifacts or other somewhat random but systematic artifacts of sequencing.

**data is encode, roadmap, geo**

**chip seq and atac seq – use macs – default cut offs are fine – you can give q value cut offs - will tell you how accessible dna is- atac seq is better for primary cell types – choose a loose cut off and run it through idr**

**methylation you can parse the file**

**for rna seq reads – use htseq counts**

**issues: data quality is bad – use cut offs – and then say statistically significant open or close**

### DNA Methylation Analysis

**Watermelon**

450K Infinium HumanMethylation450 BeadChip microarray

The standard index of DNA methylation at any specific CpG site is β = M/(M + U + 100) where M and U are methylated and unmethylated signal intensities, respectively.

Betas (βs) calculated from raw signal intensities (the default GenomeStudio behavior) perform well, but using 11 methylomic datasets we demonstrate that quantile normalization methods produce marked improvement, even in highly consistent data, by all three metrics. The commonly used procedure of normalizing betas is inferior to the separate normalization of M and U, and it is also advantageous to normalize Type I and Type II assays separately. More elaborate manipulation of quantiles proves to be counterproductive.

QN is a nonlinear transformation that replaces each intensity score with the mean of the features with the same rank from each array. It is guaranteed to produce identical array-wide distributions from any data, but whether this can be achieved without losing information depends on whether the raw distributions are suitable.

 We have taken advantage of known DNA methylation patterns associated with genomic imprinting and X-chromosome inactivation (XCI), in addition to the performance of SNP genotyping assays present on the array, to derive three independent metrics which we use to test alternative schemes of correction and normalization.

DNA methylation profiles were obtained for 138 tumors and 4 normal samples. Data were preprocessed using the “minfi” ([**73**](http://cancerdiscovery.aacrjournals.org/content/7/10/1116.long#ref-73)) and “wateRmelon” ([**74**](http://cancerdiscovery.aacrjournals.org/content/7/10/1116.long#ref-74)) R packages. We selected 4,520 probes with the highest standard deviations in β-values across the tumors, and mean β < 0.5 in the normal samples, for clustering using the “RPMM” R package ([**75**](http://cancerdiscovery.aacrjournals.org/content/7/10/1116.long#ref-75)). In the hypermethylated methylation clusters (1 and 4), we considered a CpG site to be hypermethylated if the following conditions held: (1) β < 0.5 in normal samples; (2) M values were significantly different in the (i) hypermethylated cluster versus (ii) the combined normal samples and the low-methylation tumors—those not in methylation cluster 1 or 4 (q < 0.05, two-sided t test); and (3) its mean β in the hypermethylated cluster minus the mean β across the normal samples and low-methylation tumors was >0.2.

To explore associations between mutation signatures and hypermethylated CpGs, we considered only mutations located within 50 bp of CpG probes that had mean β < 0.5 in normal samples. In each tumor, the nearest CpG probe to a mutation was considered to be hypermethylated if (i) it was hypermethylated in that tumor's methylation cluster; (ii) its individual β was > 0.5; and (iii) its individual β minus the mean β across the normal samples and low-methylation tumors was >0.2. Other analyses were similar to community-standard analyses. Further details are provided in the Supplementary Methods.

gene sets - cisbp

promoters – integrate gencode, tried hgnc, ucsc

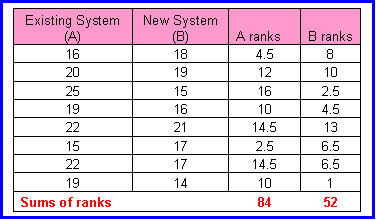
enhancers – fantom

mutation prediction – 5 method integration is best

We aligned sequence data to the human reference genome (hs37d5) using BWA-MEM v0.7.9a ([**54**](http://cancerdiscovery.aacrjournals.org/content/7/10/1116.long#ref-54)). We removed PCR duplicates using SAMTools ([**55**](http://cancerdiscovery.aacrjournals.org/content/7/10/1116.long#ref-55)) and performed indel realignments using Genome Analysis Toolkit v1.0 (GATK; ref. [**56**](http://cancerdiscovery.aacrjournals.org/content/7/10/1116.long#ref-56)). We used realigned data as input to both GATK Unified Genotyper and MuTect to call sSNVs. We used GATK IndelGenotyperV2 to identify indels. We applied filters and manual inspection to retain only high confidence sSNVs and indels. Further details are provided in the Supplementary Methods.

missing new – Wilcoxon?

Why mann whitney:



I use Wilcoxon rank sum test

**Ordinal data**

The Mann–Whitney *U* test remains the logical choice when the data are [ordinal](https://en.wikipedia.org/wiki/Level_of_measurement#Ordinal_scale) but not interval scaled, so that the spacing between adjacent values cannot be assumed to be constant.

An *ordinal* variable is a categorical variable for which the possible values are ordered. Ordinal variables can be considered “in between” categorical and quantitative variables.  
  
*Example*: Educational level might be categorized as  
    1: Elementary school education  
    2: High school graduate  
    3: Some college  
    4: College graduate  
    5: Graduate degree

•    In this example (and for many ordinal variables), *the quantitative differences between the categories are uneven, even though the differences between the labels are the same*.

**Robustness**

As it compares the sums of ranks,[[12]](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#cite_note-Motulsky_2007-12) the Mann–Whitney *U* test is less likely than the *t*-test to spuriously indicate significance because of the presence of [outliers](https://en.wikipedia.org/wiki/Outlier) i.e., the Mann–Whitney *U* test is more [robust](https://en.wikipedia.org/wiki/Robust_statistics).[[*clarification needed*](https://en.wikipedia.org/wiki/Wikipedia:Please_clarify)][[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]

**Efficiency**

When normality holds, the Mann–Whitney *U* test has an (asymptotic) [efficiency](https://en.wikipedia.org/wiki/Efficiency_(statistics)) of 3/*π* or about 0.95 when compared to the *t*-test.[[13]](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#cite_note-Lehmann_1999-13) For distributions sufficiently far from normal and for sufficiently large sample sizes, the Mann–Whitney *U* test is considerably more efficient than the *t*.[[14]](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#cite_note-Conover_1980-14)

Wilcoxon rank sum test is a nonparametric alternative to univariate logistic regression.

Wilcoxon rank sum test assumes that predictor is nominal or ordinal in nature and that samples in groups A and B, that are being compared, are independent.

t-test is for normal distributions

[Shapiro–Wilk test](https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test) **and the**[**Anderson–Darling test**](https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test) are two tests considered more powerful than the KS Test. There is a major downside with these two tests, they don’t allow you to compare two samples, they always compare a sample with a standard distribution

In particular, the Kolmogorov-Smirnov test is weak in cases when the sample empirical cumulative distribution functions do not deviate strongly even though the samples are from different distributions. For instance, the Kolmogorov-Smirnov test is most sensitive to discrepency near the median of the samples because this is where differences in the graph are most likely to be large. It is less strong near the tails because the cumulative distribution functions will both be near 0 or 1 and the difference between them less pronounced

The Chi-squared test is also used for testing whether samples are from the same distribution but this is done with a binning discretization of the data – need 80% of the “expected” values to be greater than 5 to get a chi squared distribution. The Kolmogorov-Smirnov test does not require this.

Permutation tests are a method that can be used to estimate any statistic which you specify such as difference in means, difference in medians, difference in CDF’s etc.

In [statistics](https://en.wikipedia.org/wiki/Statistics), **Fisher's method**,[[1]](https://en.wikipedia.org/wiki/Fisher%27s_method#cite_note-1)[[2]](https://en.wikipedia.org/wiki/Fisher%27s_method#cite_note-2) also known as **Fisher's combined probability test**, is a technique for [data fusion](https://en.wikipedia.org/wiki/Data_fusion) or "[meta-analysis](https://en.wikipedia.org/wiki/Meta-analysis)" (analysis of analyses). It was developed by and named for [Ronald Fisher](https://en.wikipedia.org/wiki/Ronald_Fisher). In its basic form, it is used to combine the results from several [independent](https://en.wikipedia.org/wiki/Statistical_independence) [tests](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing)bearing upon the same overall [hypothesis](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing)

X2 = -2(sum(ln((-pi)))

Ln(0.002) = -6

Ln(0.002) = -3

 When the p-values tend to be small, the test statistic *X*2 will be large, which suggests that the null hypotheses are not true for every test.

When all the null hypotheses are true, and the *pi* (or their corresponding test statistics) are independent, *X*2 has a [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution) with 2*k* [degrees of freedom](https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)), where *k* is the number of [tests](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing) being combined. This fact can be used to determine the [p-value](https://en.wikipedia.org/wiki/P-value) for *X*2.

in addition to this is also worked on – and am looking forward to translating that to cvmd

* how is your research related to what we do
* how to talk about your post doc project

I am a Postdoctoral Research Associate at the Duke University School of Medicine, where I lead a team of computational biologists, statisticians and bioinformaticians to identify novel, rare biomarkers in cancer. In a recent high impact project, I was recruited to a consortium of 55 scientists located across 8 countries to identify noncoding biomarkers in bile duct cancer. A key challenge with this large-scale project was in adapting to rapidly changing project goals and timelines. The noncoding analysis evolved from having a supporting role …to validate existing role….because we got such good results…to the main focus of the project. This required a substantially expanded set of analyses to be completed in the original time frame. In order to adapt to this new goal, I quickly developed a new project execution plan, interviewed and trained additional bioinformaticians, structured the project – see who does what in what time frame – I trained them - enlisted the cooperation of Duke University experimentalists to validate our findings, organized regular meetings with our collaborators to facilitate any mid-course changes, and motivated the team to complete their tasks ahead of schedule. On the technical side, I designed and implemented high-speed, big data pipelines that integrate diverse sources of data including next generation DNA sequencing data, RNA expression data, epigenetic information and protein-DNA binding data. I statistically analyzed the data using Supervised, Unsupervised Learning and Principal Component Analysis (PCA), conducted gene expression analysis using DESeq2 and network/pathway analysis using GSEA, GSA, IPA and STRING. The project resulted in the breakthrough discovery of noncoding biomarkers associated with bile duct cancer. The findings were published in the high impact factor journal, Cancer Discovery.

Single nucleotide mutations play a major role in the onset and proliferation of cancer. Mutations in coding regions have been comprehensively analyzed. Nevertheless, even some well-studied cancers do not have known driver mutations in major subpopulations. On the contrary, the impact of non-coding mutations, which recent studies suggest constitute 85-90% of disease-associated mutations, remains widely unexplored

We applied a preliminary version of our method to cholangiocarcinoma whole-genome sequencing data and we found significant dysregulation of epigenetically important H3K27me3- associated regions (Jusakul et al. *Cancer Discovery* 2017). We further found that transcription factors that regulate the epigenetic machinery are dysregulated in the tumor DNA compared to the normal DNA. These findings were consistent with gene expression data derived from primary tumors, histone modification and DNA hypermethylation analyses, and suggest that epigenetic modification plays a pivotal role in cholangiocarcinoma.

Regulatory region mutations – 85-90% of disease causing mutations- given the current wgs and high throughput technologies – we can finally analyze them

What we get out of that is:

1. you can use regulatory regions to identify genes pathways and tfs in cv disease- and also help select patients for clinical trial - can be different for patients from different parts - so this is where precision medicine comes in

-this can lead to different drug targets -therapies with maximal efficacy and no or limited adverse reactions

1. - dysregulated tfs are seen in cv disease - transcription factors have been used as drug targets- now though with wgs data and also high throughput assays that inform how tfs work - we can informaiton that was not possible before
2. - this can be used to treat even before symptoms emerge

**Assumptions being made: include in conclusion**

1. sequences on either side of the mutation are the same as the reference
2. not necessarily filtering out the effects of cnv’s, structural variants, etc –

|  |  |  |
| --- | --- | --- |
| substitutions | insertion/deletions | rearrangements |

1. duplications
2. epigenetics?
3. should we be grouping patients into subtypes
4. Should we do a reverse approach, i.e., look at genes with changes in expression (even small changes) and identify “significant” mutations in their putative promoter and enhancer regions. If the results from both approaches overlap, we can have higher confidence that we are not just picking up noise. This approach works better for the TCGA data sets where we have expression data for all patients.- dont have enough matched data
5. bile duct – doras pipleline is using the ref genome as “normal” for sequences – so if the mutation in the tumor is the same as the normal – in my prediction script, these mutations i ignore
6. not taking germline mutations into account
7. pbm score cut offs
8. promoter defined as +/- 2kb?
9. validation that tf binding changes – gene and pathway expression changes

1. chip seq bias – gc bias in cross link, efficiency of antibody, size of fragmented dna, pcr bias, gc bias in seq
2. pindel has 80% indel prediction accuracy
3. for masking – not taking all of the masking programs criteria – like in a 50 bp window if more than 30% of calls are filtered, we don’t take that into account

The ideal way to do this:

single cell – identify the original cluster of cells that “cause” the cancer

– take tumor and normal from the tissue – a few of these

– do all analysis ie sequencing, accessibility, protein expression –high quality – time points

- then identify genes/pathways/networks that are disrupted

to improve signal to noise computationally:

1. use signals levels in individual assays
2. use lots of controls – with anything you are not sure of

phd

my phd project was to identify rna biomarkers specific to human cells

– many diseases have aberrant rna secondary structure – including cv diseases – aberrant secondary structure in the braveheart non coding rna

there is no method that can probe rna structure in a live human being – – detect and potentially treat disease rna biomarkers

I developed a biosensor that can assemble on structural biomarkers – and based on the images generated from the biosensor – you can say if the biomarker is from a tumor or normal

this is the first– novelty led to best research poster and paper

both computational and experimental – computationally - for this I used a baysein approach – develop a ctmc

flouorophore labeled dna strands that bind to single stranded regions in the rna structure – then I use the model to see if the structure can be differentiated from the normal structure

* stochastic processes are defined as a set of random variables indexed in time – state space was discrete
* transition rate is independent of time, it is referred to as a homogenous process
* Transfer rate given by FÖrster’s equation:
* Ret is stochastic processes - Fluorophores modeled as random variables with predefined probabilities of transfer
* Memoryless
* The time that a homogenous, continuous time Markov chain spends in a given state is known to exhibit an exponential behavior similar to that of the time spent in the excited state of a fluorophore
* Emission was modeled using an absorbing state
* Dx/dt(1) = -(a12+a13+a14+a17)\*x(1) + a21\*x(2) + a31\*x(3)
* 
* dx/dt(2) = a12\*x(1) - (a21+a23+a25+a28)\*x(2) + a32\*x(3)
* implementation was a bunch of differential equations – for each fluorophore or RV capture all the incoming transitions into a state- and also outgoing transitions – do that for all the fluors in the network– numerically solve in matlab

– also information theoretic approach to model rna secondary structure in disease states –

to figure out how large we can make these networks

–cause there are only so many fluorophores – and when you measure the output you have to get a measurement from all of them – the fluorophores that you are changing and the others that are not being changed

upto a 100, 200 – but youll get to the point where you add one more fluorophore – in change in signal may not be that much- so created an information theoretic model to see p log p of the trasitions in the network – at what point will it saturate

also validated the sensor using in vitro experiemnts – novelty led to best research poster and paper

this might translate

the image analysis

– you get a fluorescence image from the structure

– use hough transform to get straight lines from the image - to multi exponential decays – apploying log - we convert that to straight lines - use a hough transform to get fit the lines and get rid of noise

* convert to a binary image – for each important area of the image
* hamming distance – same samples measured different times tell you how much diff in hamming distance is alright
* tried not converting image to straight lines – tried using correlation – tried using amplitude and lifetime of exponential decays – degree and distance of the hough transformed lines

also medicines of tomorrow – using oligonucleotide conjugates, antisense oligonucleotides

if you can detect this, you can detect cardiovascular disease early –treatment - A RNA (ribozyme) or DNA (deoxyribozyme) molecule that exhibits catalytic properties to cleave RNA in a site-specific manner, Chemical compounds that block the activity of target lncRNAs by structure-specific docking

Among biologically validated lncRNAs, several have been associated with cardiac development ([Table 2](https://www.sciencedirect.com/science/article/pii/S2162253117302251" \l "tbl2)). For example, Braveheart (Bvht) has a critical role in cardiac lineage commitment in mouse. It is abundantly expressed in embryonic stem cells and regulates the transition from nascent [mesoderm](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/mesoderm) to cardiac progenitor.[24](https://www.sciencedirect.com/science/article/pii/S2162253117302251" \l "bib24)Bvht, by modulating the core cardiovascular [gene network](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/gene-regulatory-network) and mediating the epigenetic regulation of cardiac fate, is necessary to maintain cardiac commitment.[25](https://www.sciencedirect.com/science/article/pii/S2162253117302251" \l "bib25) Conversely, the lateral mesoderm-specific lncRNA Fendrr(fetal-lethal non-coding developmental regulatory RNA) controls mesodermal differentiation, as well as heart and body wall development, by binding to the histone-remodeling [polycomb](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/polycomb-group-proteins) repressive complex [PRC2](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/prc2) and TrxG/MLL to modulate [chromatin](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/chromatin) status.[26](https://www.sciencedirect.com/science/article/pii/S2162253117302251" \l "bib26)

BvhtdAGIL EBs also displayed a failure to activate genes associated with the cardiac contractile apparatus such as cardiac [troponin T](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/troponin-t) (cTnT) and [myosin heavy chain](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/myosin) genes, whereas bvhtH4mis EB



I was involved in an international project across 8 countries – to analyze the wgs study of bile duct cancer – through the framework I developed we discovered pathways and tfs that are dysregulated in bile duct cancer– these were experimentally validated using some studies – published in cancer discovery

Also include accessibility information

Biomarker data integration

Then Im just wrapping up a project where I extended the analysis to 9 diffferent cancer types for which we have wgs data

I’m always very excited to discover new targets that can lead to therapies for patients – so I applied to this job to see if the 85-90% of the noncoding regions can be used to discover new genes/pathways associated with cvmd – I haven’t seen this done yet in the literature

For my phd one of the projects I worked on was to identify RNA based structural biomarkers in breast and lung cancer….for this I used both baysein and information theoretic models

Merck

* tranditionally you have a single biomarker or target for most diseases
* but many diseases and response to therapy is more complex - but genomics, proteomics, metalobolomics, microbiome, structural imaging features that you get from an ultrasound
* biologically figuring out the connection between say genomics and transcriptomics is unsolved and will take decades
* the idea and this what the merck innovation award was for – it was a 3 part proposal – one part was data integration
* use rf for feature selection – then nn – to determine which combination of features – use gradient descent with stochastic or baysein optimization
* gradient descent – don’t know if we’ll get to global minima
* a priori algorithm – what if we have unequal amount of data for each data type

1. Which biomarkers are relevant for the disease – also more accurately subtype- choose patients for clinical trials – reduced toxicity, improved efficacy

2. we could use to identify combination of targets for therapy

* cant do linear regression – cause there might be a nonlinear relationship between input predictors and outcome
* svm – needs scores for different classes and it will classify – but then how do we know how to get a score for different input feature combinations – we don’t know which is important and by how much
* naïve bayes – input predictors not necessarily independent – also you need to have the input predictor values binned and you may not know the binning
* random forests – again need to know how to split continuous variables – but otherwise its good cause it tells you which variable combinations are important and how much - will not get stuck in local minima
* nn
* use gradient descent – features increase in complexity –think genes to pathways – pathways value updated with one set of training data – then a gene updated with another set – stochastic gradient descent so it doesn’t get stuck in local minima – will though if cost function is nt convex
* – good cause it can capture info about which variable combinations are important and how much – can have a nonlinear relationship – maybe captures dependence better than rf because its not greedy?

Firstly, we can classify personalized medicine into comprehensive personalized medicine (e.g., family history checking, classic risk factor assessment, laboratory testing, tailored medicine) and more detailed clinical personalized medicine as a narrow meaning (e.g., only drug therapies through genetic testing), conceptually. Secondly, we can classify from genomics to proteonomics, by the expression path of genes (e.g., genomics, transcriptomics, proteomics, metabolomics). Thirdly, we can divide the structural {e.g., intima-media thickness (IMT), intravascular ultrasound, optical coherence tomography} and functional personalized medicine (e.g., endothelial function, exercise testing, heart rate variability etc.). Lastly, we can classify personaliz2ed medicine according to clinical objectives, applications, and therapie

astrazeneca

* what do you know about cardiovascular and metabolic diseases
* need to figure out what your ideas for cv md ar
* 5 year plan

stay in computational biology – eventually lead a team in the area – its actually something I do now and enjoy very much? – look into chanchals profile

I really enjoy comp bio - Ive all these ideas

it might be ambitious – but I really like working in this area – and I am by nature very hardworking

maybe be more specific? – no cause it may not align with what they want

* why do you want to work in this company

I developed this statistical method to identify the regulatory contribution to a pan cancer –we found very interesting results – bile duct – validated experimentally – now also breast and lung cancer which we are testing in the lab - I want to now extend it to other diseases

In the process of transitioning from a post doc in academia to industry – I looked at a lot of companies – Astrazeneca is definitely the most exciting for genomics/personalized medicine/cancer –90% of pipeline is dedicated to personalized medicine – 6 new cancer drugs to the market by 2020 - its goal to analyze 2 million genomes by 2026 brings up very exciting opportunities for someone with computational and statistical genomics experience

1bn in research – still publish a lot

* can you relocate

Yes definitely. Im not married and don’t have children so at this point I have complete geographic flexibility

My fiancée and me want to relocate to Europe cause its closer to home

I should be able to wrap things up here in a month

Why Sweden:

Sweden is also voted as one of the best places to stay – and Ill be closer to family

If you don’t know something – say I would like to learn about it – ah I need to read up on that – I know someone who works on that so I can learn before I join

* multidisciplinary team to achieve team objectives.
* I was recently on a project which was the first wgs study in the world – it was a collaboration across 8 countries and over 50 scientists

Merck team – from different continents

* Also in lab Im constantly surrounded by people who are from different continents
* I also really like the diversity – have had roommates from 8 different countries so far!

The challenge:

Time: once we saw that the noncoding region analyses started giving good results – most everyone else wanted to cross check the results with the results from their analyses – so a lot of people were asking for a lot of things – 5 different people would ask for 5 different analysis at the same 2 hour window - quickly I realized that I need to communicate – so I started saying im doing this for this person – can I give this to you at this time – or if it was important – I would email everyone concerned and ask if we can reprioritize

Culturally –here in the US, we are encouraged to think independently and point out issues – some countries are hierarchical where you do what the pi says – so I learnt how to point things out – not to do it in front of others – and also instead of saying this is an issue - ask could this be a concern and let them respond

Multivariate analysis:

* pca, hierarchical clustering, regression to see multiple time domain fluorescence signals how they compare with each other

Can you learn things fast

Im a quick learning – I really like being good at what I do – if that means I need to work 7 days a week to learn something new ill do it – that is reflected in the schooling – topped university in undergrad – moved to nanoscale physics – even though that was the first time I was doing physics – A+ - phd

– then I decided I really like applying computational and statistical methods to help develop therapies for patiets so I stayed in that area for my post doc

4 years when the average graduation time for experimentalists in my lab is 6 years – 2 breakthrough projects –  paper in best journal in the field – impact factor 15 -invited to speak at multiple conferences including the biggest in our field -paper in nature journal – also won the best research poster award

i took 4 classes even though i finish my course requirements –spoke to people in the field -2 years of working 7 days a week later,

why cant we just do this with a protein analysis or rna expression:

* you’ll know this protein is dyregulated – but you wont know why - need to find targets – you’ll know because of this mutation – this tf is dysregulated – so this gene is affected – you can target that – do we know which protein how much should be in each cell – even if we know there is a cascade and we cannot attack the root cause leading to less efficacy and greater toxicity
* lncs rnas turn on and off genes – so involved in epigenetic control – so this is something analyzing genomics/transcriptomics might help

problems:

-quality of data

- heterogenoues – so power issues – subtyping – improves with quality

how can you be confident of the results – no fps?

* validate with experiemental data of same study
* validate in independent studies

**by heart answers**

* questions for us

1. gmd - and what your goals are for the new future- What are the ideal qualities in the candidate you are looking for –qualities that help you succeed at AZ - what do you like about working here
2. Could you please talk about your team – and what your goals are for the new future
3. - What are the ideal qualities in the candidate you are looking for
4. - what is the culture like
5. how is Sweden to live
6. What do you like about astrazeneca
7. - if you want to learn new things to develop your career, do they support
8. I always try to figure out qualities of top performers in this field – you finished your phd in 2008 and you are already the associate director of stats and machine learning

– also worked in roche and Novartis – how do you like astrazeneca

Specifically for you – what would you say led to the quick growth from the time of your phd to heading the cvmd translational science dept

In a short amount of time – you achieved an extraordinary amount -  Prolific publication record – both number and citations – to associate director –being able to lead interdisciplinary projects - I would very much like my career to follow along the same path – could you talk about how you did it and do you have advice for someone who’s just transitioning to industry

For the team – is it collaborative, how is living in sweden

Later – if I get in what can I do to be prepared

**Links:**

**Position:** [**https://job-search.astrazeneca.com/job/molndal/combined-biostatistics-and-bioinformatics-expert/7684/6197770?utm\_source=jobalert&utm\_medium=email&utm\_campaign=GlobalTB\_jobalert&utm\_content=R-016500&ss=paid**](https://job-search.astrazeneca.com/job/molndal/combined-biostatistics-and-bioinformatics-expert/7684/6197770?utm_source=jobalert&utm_medium=email&utm_campaign=GlobalTB_jobalert&utm_content=R-016500&ss=paid)

**http://www.euraxess.it/jobs/305391**

**Metabolism:** [**https://www.cell.com/cell-metabolism/fulltext/S1550-4131(17)30092-X**](https://www.cell.com/cell-metabolism/fulltext/S1550-4131(17)30092-X)

**Phenotyping:** [**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4133145/**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4133145/)

**Noncoding rna:** [**https://www.sciencedirect.com/science/article/pii/S2162253117302251**](https://www.sciencedirect.com/science/article/pii/S2162253117302251)

**Biomarkers:** [**file:///Users/vishwanellore/Downloads/1666\_Cirion\_factsheet\_biomarkers\_V3\_web.pdf**](file:///Users/vishwanellore/Downloads/1666_Cirion_factsheet_biomarkers_V3_web.pdf)

**Personalized medicine in cv:**

**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467440/**

look up gene identifier mapping – from this site – look up different identifiers – also can webscrape

look up geo – snp, expression, methlation raw data for different stdies

The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

compare transcriptomic profiles

* could normalize to housekeeping genes
* The Z-score test was applied independently on micro-array and single-cell RNA sequences for profile comparison.

In meta-analysis, multiple studies are analysed separately and statistical results are subsequently integrated. By contrast, merging means combination of multiple comparable data sets into one big data set prior to statistical analysis

<https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-017-10930-w/MediaObjects/41598_2017_10930_MOESM1_ESM.pdf>

We use multiple end-points (fibrosis, steatosis, inflammation, NASH vs SS etc). We use linear or logistic regression models to test for association between the end-points and mRNA levels. Statistical results are regression coefficients and associated pvalues. We integrate these statistical results to obtain the final gene signature in two steps

First, we identify end-points that are present in min. 3 data sets. We use such end-points to obtain intermediate gene lists (list of genes associated with fibrosis, list of genes associated with inflammation etc). Second, we merge these intermediate 2 gene lists and obtain gene signature of NAFLD progression.

* Dony want to do this in our case cause we want to aggregate counts across pathways
* Also diff patients may have mutations in diff genes

Models were adjusted for the most likely sources of variation (e.g., BMI) when the information was publicly available and sample size permitted estimation of a multivariate model?

Spearman correlation coefficients were computed for all pairs of probe sets representing genes in NAFLD progression signature and genes with genetic evidence for NAFLD.

Papers from cvrm group:

<https://www.sciencedirect.com/science/article/pii/S0085253818303569#mmc1>

<https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-017-10930-w/MediaObjects/41598_2017_10930_MOESM1_ESM.pdf>

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019095#authcontrib>

<https://job-search.astrazeneca.com/job/cambridge/clinical-information-science-principal-oncology-and-immuno-oncology/7684/8939258>

**Senior Data Scientist, late stage clinical trials**

* Tell me about yourself

Im currently a post in the duke school of med where I lead a small team to identify genetic features correlated with cancer patients – while most people look at

For this I integrated diverse sources of high dimensional data to identify which genes are associated with the disease – skills I used were unsupervised learning with pca – supervised learning to classify transcription factors to a family – network analysis to identify which gene networks are associated with the disease – as well as stat techniques fishers and Wilcoxon mann whitney to compare two distributions

For my phd I developed a biosensor that can distinguish between different structural biomarkers in cancer through image analysis – there is no method currently to identify structural biomarkers in cancer

For this I used baysein methods specifically ctmc – and information theoretic models so entropy modeling - to create the sensor itself – and then used image analysis methods - hough transform, pearson correlation – Wilcoxon rank sum

At merck – I recently got the innovation award for a 3 part proposal – one part was to integrate different biomarkers –

For this I used – web scraping and text mining to get information from clinical trial databases – rf for feature selection – nn and apriori algorithm to be implemented

**Clinical Information Science Principal, Oncology & Immuno-Oncology**

* Tell me about yourself

Ill start off by explaining my experience relevant to this position

My experience in my phd and post doc is in identifying biomarkers to help select patients for 1q AZclinical trials - both oncology and immuno oncology

For my post doc – I conceived and implementd a project was to subtype patients with bile duct cancer – for this I developed novel bioinformatics pipelines to identify genes affected in different sub groups of patients– these were very novel results that was published in cancer discovery

This can be used to subtype patients for clinical trials

For my phd – my project was to identify structural biomarkers in lung and breast cancer cells – for this I developed novel machine learning and information theoretic models – and the novelty of this work led it being awarded the best paper in 2017

So while the work in academia was oncology related

I got the merck innovation award for an immuno oncology project idea –– where I conceived and developed a method by which we can integrate multiple types of biomarkers to subtype patients who will respond best to cancer immunotherapy – for this idea, I actually got to go to merck this summer and implement the idea from a tech and business perspective – and our team also won the vibrant digital innovation award

Subtyping patients for clinical trials is the big area of research in all pharma companies – and that’s what my research was focused on

In addition I also know how to do all the different standardized statistical analyses in late stage clinical trials including sample size estimation, randomization and endpoint determination

az

I got my phd from duke in comp engg where I learnt many of the computational and statistical skills – and in my post doc I focused on applying those computational and statistical skills to identify biomarkers for patient stratification in oncology and immuno oncology

With respect to clinical information science I have experience in finding novel sequence and structural biomarkers in oncology and immune oncology

* Sound understanding of the pharmaceutical clinical drug development process (with a preference for late-stage drug development in oncology, immuno-oncology of immunology)

Training –

I had classes on clinical trial design and statistics for clinical trials

Different study designs, randomization, picking patients, analyzing data

Statistics for clinical trials – where once we saw the data -

Exprerience –

Ts fellow at NIH – Where I got to see how drug development happens from start to finish including phase 2 and 3 clinical trials, fda review

* Trained in drug discovery and development from experts in translational science; topics included target identification and validation, intellectual property and the patent process, pre-clinical and clinical trial development, FDA review, marketing and drug product launch.





and Merck – was developing a method for cancer immunotherapy patient selection in late stage clinical trials – the technical component – mining clinical trial data to see what biomarkers are there- then doing machine learning to figure out if the biomarkers are correlated with patients who respond to treatment - and this included the whole identifying competitor landscape, patent issues and financial viability of the product.

All this in addition to my job for the last 7 years - phd and post doc – developing biomarkers that associate a patient with a cancer

So target identification – all the way to fda review- Ive had several levels of experience

* Good understanding of project management techniques and methods
* Conceived, designed and implemented a project to understand the molecular mechanisms across 11 cancer types to identify new biomarkers and help clinicians with cancer diagnosis, progression and treatment.
  + Developed statistical models and data analysis pipelines that integrate multi-omics data including next generation whole genome sequencing data, RNA expression data and epigenetic information; functionally characterized genetic variants;
  + Communicated with 55 scientists across 8 countries on a weekly basis regarding project goals, coordinated project execution strategy and ensured the timely completion of the project module.
  + Identified regulatory regions, pathways and transcriptional mechanisms dysregulated in bile duct cancer.
  + Resulted in the breakthrough discovery of the molecular mechanisms involved in first WGS study of bile duct cancer.
  + Findings published the high impact journal, Cancer Discovery.
* I am a Postdoctoral Research Associate at the Duke University School of Medicine, where I lead a team of computational biologists, statisticians and bioinformaticians to identify novel, rare biomarkers in cancer. In a recent high impact project, I was recruited to a consortium of 55 scientists located across 8 countries - different cultures, languages, time zones -  to identify noncoding biomarkers in bile duct cancer. A key challenge with this large-scale project was in adapting to rapidly changing project goals and timelines. The noncoding analysis evolved from having a supporting role to the main focus of the project. This required a substantially expanded set of analyses to be completed in the original time frame. In order to adapt to this new goal, I quickly developed a new project execution plan, interviewed and trained additional bioinformaticians, enlisted the cooperation of Duke University experimentalists to validate our findings, organized regular meetings with our collaborators to facilitate any mid-course changes, and motivated the team to complete their tasks ahead of schedule. On the technical side, I designed and implemented high-speed, big data pipelines that integrate diverse sources of data including next generation DNA sequencing data, RNA expression data, epigenetic information and protein-DNA binding data. I statistically analyzed the data using Supervised, Unsupervised Learning and Principal Component Analysis (PCA), conducted gene expression analysis using DESeq2 and network/pathway analysis using GSEA, GSA, IPA and STRING. The project resulted in the breakthrough discovery of noncoding biomarkers associated with bile duct cancer. The findings were published in the high impact factor journal, Cancer Discovery.
* Hired computational biologists, statisticians and bioinformaticians, designed computational and statistical solutions to analyze mutations in cancer, structured the project into sub-tasks, trained the team on statistical and biological concepts key to the project, assigned responsibilities and timelines, removed process inefficiencies, managed change in project scope, guided the team’s efforts on a daily basis and provided performance feedback.
* I lead a group of 3….one of the kept sticking to the methods she knew even if they were not ideal in that situation –Giving negative feedback is difficult for me but in this case it had to be done - and I try to align my goals with the goals of the team members – I knew she was looking for jobs in industry soon – so I said that this is a skill that is often used in genomics so having this skill set will bode well for you when you do interviews – so then she took it on and learnt

now in my role as a post doc, I lead a team of statisticians, comp bio and bioinformaticians to identify oncology biomarkers that can be used to identify patients for clinical trials – have extensive and very diverse project management experiences over the last 10 years including project manager and president

or say as project manager this is what I did – managed so many people - result

* Experience in clinical information or science information setting in the application of information and knowledge management (settings could include, but are not limited to clinical drug development, medical writing, medical affairs, clinical operations, competitor intelligence, regulatory intelligence)

See text mining section

Academia – this the problem – go through the current literature with pubmed, etc to figure out which information can be used to find a solution

That experience really helped – at merck – where we needed to find out how best to select patients for cancer immunotherapy clinical trials - in figuring out what information we need to come up with a solution – also saw what the competitors are doing by going through the clinical trial literature and patent info – and also searched through regulatory intelligence to see if we can get approval

* Good organizational skills and the ability to multitask; can set priorities and follow a timeline

Been my career so far actually – multitasking and finishing tasks to a deadline

Phd –most people do 1 but I juggled 2 projects – did both computational and experimental to improve the strength of the results – got published in high impact journals and also won best paper award – but I managed time efficiently and finished in 4 years while the avg graduation time in my lab is 6 years

Post doc – even though I implemented all the different parts of my project – finished in under 2 years – usually in much bigger teams, it takes longer

* Great attention to detail
* When im reviewing literature I make sure I check all the fine print and all the assumptions being made
* When I implement methods I short list methods that fit – then I make sure that I satisfy all the criteria for a method

Clinical information science principal

The position

This is a unique position – its not analysis of clinical data – its getting information that is useful for the analysis – could involve reuse of existing data

Lead a team of 12 clinical directors & principals, plus industry collaborations, partners and suppliers   
driving re-use of clinical data for Oncology & Immuno-Oncology drug development  
Development & execution of the Oncology & Immuno-Oncology Information Strategy to maximize value across the portfolio  
Member of Information Practice Leadership team and the B&I Oncology & Immuno-Oncology Leadership team

Been creative and unconventional in research – merck, post doc, phd - best paper

2 things:

* What info to look for
* Creative/unconventional
* Done clinical trials/oncology
* How will you look for it – web scrape, text mining which uses a lot of machine learning – for standard and non-standard data, statistical analysis, visualization
* Good results
* Soeone who look at clinical trial data – given this trial – in other data, what were the number of patients used, randomization, study design, this drug worked – statistical analysis of how important something is

What I should bring up about my background:

Just that I have a broad range of skills by training – and that was good vause rgis was actually 100 diff analysis – and I used

* Given that this position requires experience with oncology/IO clinical trial data – both to identify the data needed, mine it from diff sources, analyze it so it can be used downstream - pretty much all of my experiences are in those areas – so I can talk about them
* – I’ll actually start with my experience at merck even though I just started and go back from there – and we just found out that our project was selected to be implemented
* I was selected from over 2200 people from around the world for the cup - merck does this innovation cup every year – shortlisted in the first round based on the CV – and then you submit a innovative, out of the box idea and if your idea is selected you get to go there and develop it further
* At merck – the project was selecting patients for clinical trials – by predicting which groups of patients respond to cancer immunotherapy – by combining biomarkers from different areas
* tranditionally you have a single biomarker or target for most diseases
* but many diseases and response to therapy is more complex - but genomics, , metalobolomics, microbiome,
* biologically figuring out the connection between say genomics and transcriptomics is unsolved and will take time
* – for this I had to use both web scraping and text mining to identify which biomarker information is available from clinical trials.gov – and figure out how many times information about a biomarker is collected
* once we had that data–I built a preliminary model that identifies the most important features using the random forests algorithm - and that information was to be analyzed neural network algorithm
* – so this won the innovation cup in the vibrant digital category - we recently found out that the project was taken up for full time

my post doc and phd was in identifying different types of biomarkers in cancer – and this can potentially be used to identify targets and also subtype patients in clinical trials - finished

I could talk about them in a couple of minutes if that’s ok – the skills I used

* For my post doc –I lead a team of biostat, comp bio and bioinformaticians- to identify sequence based biomarkers in cancer – genes/pathways/tfs - and this can be used to subtype patients say for trials and also identify targets to develop therapies
* I took an unconventional approach- while most people look at coding regions – we still don’t know the cause of most cancers- and looked at noncoding regions of the genome
* For my post doc, I conceived and implemented a project that focuses on analyzing mutations in the noncoding regions of the genome so promoters and enhancers - we know that even though coding reions are analyzed – we still don’t know the cause of most cancers
* – again the literature search was in journal databases, which is more unstructured data compared to the structured data I used at merck –
* epic project across 11 cancers – the goal was to find sig genes/pathways/tfs - integrated wgs, rna, dna-prtein binding data - and used statistical analysis methods from a broad range of classes that I took – I used most of them here– which resulted in me finding genes that were known before as well as new genes – published in cancer discovery
* I was recently involved in an international collaboration between 8 countries to analyze the first wgs study in bile duct cancer – we discovered pathways that were dysregulated in bile duct cancer– these results were also experimentally validated – and the preliminary study was published in cancer discovery. - but we didn’t find it by analyzing mutations in the coding regions alone
* Finally for my phd –
* my phd I conceived a project to identify rna structural biomarkers in cancer cells
* for this I used prob theory, baysein stats with a ctmc, information theory with an entropy model
* and validated the model using experiments – dna self assembly, fluorescence microscopy, afm
* this with my advisor away
* – many diseases have aberrant rna secondary structure
* there is no method that can probe rna structure in a live human being – – detect and potentially treat disease rna biomarkers – cause bio cant model florophore networks – I have cs background – can specifically burn the cancer structure
* this is the first– novelty led to best research poster and paper – nanodds, bibe last year
* I took an unconventional approach – by combining my bio and comp training - I developed a biosensor that can assemble on structural biomarkers – and based on the images generated from the biosensor – you can say if the biomarker is from a tumor or normal
* A fundamental problem in current cancer therapies including chemotherapy is that both cancerous and normal cells are eliminated leading to severe side affects in patients - as a computer engineer when i heard that certain breast cancer cells had a different dna structure compared to normal cells - i had an idea - what if using biological elements, we could make a network around the cancerous structure - it'll be an optical network - when you shine a laser pulse from outside, based on the optical network formed you'll get a different color light out - but in the normal cell, the network won't form and when you shine a laser pulse, you'll get the same color light out - to implement this though - see challenge
* I had to learn about CS, graph theory also DNA, biology, chemistry - computational modeling techniques - experimental techniques - in a short amount of time - took classes, spoke to experts, went to conferences - i had to do this all by myself - my advisor was on sabbbatical - when he was back -  i finished the paper - presented at the nanomedicine conference 2 years ago - won best research award - i completed my phd in record time of 4 years even though it typically takes 6 years for people like me doing experiments in my lab

moving forward my dream job is to use my skills to help develop therapies for cancer patients

Creativity – came from wide range of coursework – or for my phd – lit search was done using web scraping and text mining

I also used web scraping to identify which proteins group together

Can use it in clinicaltrials.gov, medline abstracts – to figure out for cancer immunotherapy what type of data is collected

Have lots of statistical analysis methods that can be used to classify

 visualizations of patient timelines, predict patient outcomes based on larger trends, and show comorbidities between different risk factors and their indicators – python seaborn cuase you can make pretty and quick plots – I use ggplot in R – and Origin – and Tableu

AZ values:

We follow the science.

– keep myself up to date – still take courses on coursera

- one of my pet peeves with stats is that – you can get statistically significant results but in reality its just from noise and it doesn’t mean anything

- that’s what I do in my project – when I run the first leg of my analysis I get a lot of significant genes that are correlated with cancer – I didn’t say oh that’s great so many results to publish – I wanted to control for as many biological confounders as possible – finally the results were down to a 10th of what we saw initially but at least they correlated well with experiemental validation

- its better to build a simple model that captures as much of the biology – than build a fancy model that just puts things together and hope you get some significance

- take classes on molecular biology and cancer

We put patients first.

When I started my phd – I was a computer engg working in cryptography – I liked the computational analysis part but someone told me that if you want to figure out what you are passionate about think if you would do it for free – not crypto – then I saw that applying my comp skills for cancer research I would do for free –that’s why I changed my project to a cancer project in my phd – continued in cancer for my post doc - I want to work in industrycause there the results have to benefit patients in the clinic

We play to win.

Over the years I've become that person - doing everything in my control to get desisted outcome is a habit - I have to force my self to not work - say Friday night go home

– undergrad – I studied everything – worked through weekends – even though post doc was in computational bio – the specific project was new to me – but my PI saw that I quickly became independent and started to lead to the project – so she said I could get my own team

We do the right thing.

* I lead a group of 3….one of the kept sticking to the methods she knew even if they were not ideal in that situation –Giving negative feedback is difficult for me but in this case it had to be done - and I try to align my goals with the goals of the team members – I knew she was looking for jobs in industry soon – so I said that this is a skill that is often used in genomics so having this skill set will bode well for you when you do interviews – so then she took it on and learnt
* Asian

We are entrepreneurial.

- during my phd when I started I was given a project –and I could’ve graduated when I finished that

- but I was excited about all the ideas I had - using my skills to cancer research – and this was all when my advisor was gone on sabbatical - I created a model of a network based sensor to detect biomarkers in cancer – experimentally validated it –using my existing skills, taking classes and talking to others at duke about things I wasn’t familiar with - that’s the one which won best paper award

- and time wise too – my advisor was super chill – but I went to lab every weekend – stayed past midnight – even though I did 2 big projects – I finish my phd in 4 yrs though the average graduation time in my lab is 6 yrs

### **Design and Interpretation of Clinical Trials**

# Understanding Clinical Research: Behind the Statistics

Clinical trials – evaluate saftety – efficacy, dosage

Late stage clinical trials:

Phase II

Since Phase I trials are usually performed in healthy volunteers, Phase II is the first stage of the drug development process, in which the new drug is tested in patients. The aim of Phase II trials is to find out if the new drug is effective in patients through hypothesis testing and at what dose the balance between efficacy and safety is optimal. The number of patients in these trials range from around thirty to a couple of hundreds, and they usually do not last longer than 2 years[[1]](http://www.cancer.net/navigating-cancer-care/how-cancer-treated/clinical-trials/phases-clinical-trials" \t "_blank). In these trials usually a new treatment is compared to placebo, sometimes also to an already existing treatment or a placebo treatment. In Phase II oncology trials all patients receive the same dose, while in other trials participants are randomly assigned to different treatment groups. These groups may get different doses or get the treatment in different ways to see which provides the best balance of safety and efficacy. In non-oncology drug development there are typically 2 types of trials in Phase II:

* Proof-of-concept trial (Phase IIa): Testing a high dose and a control, often placebo
* Dose-Finding trials (Phase IIb): If the proof-of-concept trial is positive a dose-finding trial is performed testing multiple doses and a control in a parallel groups setting.

If the new treatment is found to be equally or more effective than the existing treatment, then the study progresses to Phase III trials, where the drug or treatment is tested on more individuals. The failure of a drug in Phase II trials mainly occurs when it is discovered that the drug has some toxic side effects that were not observed in Phase I or failed to show sufficient efficacy for the medical condition under question.

Adaptive and two-stage designs, which allow for early decision-making, are attractive in the Phase II setting, because they can serve the interests of the patients as well as the drug developer. Trials with ineffective or possibly toxic drugs can be stopped early, which reduces the risk for the patients. Additionally, adaptive designs allow for changes to an ongoing trial, for example by changing the dose levels, which are used in the trial. Since dose-finding plays such an important role in Phase II, designs which allow for effective dose-response modelling are also commonly used. This usually means choosing dose levels for the different treatment arms in a way that allows meaningful information about the relationship between dose and response to be obtained.

Modelling plays a crucial role in Phase II. Nonlinear regression, mixed, hierarchical and pharmacological models are useful tools, especially if one is interested in modelling dose-response relationships. MCPMod**.** is a relatively new method, which can be used both for trial design and analysis in dose-finding trials and is therefore a very important analysis method in Phase II. In addition, interim analyses, which are performed before the completion of the trial, are an important tool. Related to these are multiple testing methods, which have to be applied when several evaluations of one or more endpoints have to be performed within the same trial.

One of the major statistical challenge in Phase II is the relatively small sample size, something which always makes a statistician’s life difficult. Another challenging aspect is the decision of choosing between a more model-based or a hypothesis testing type of approach. Since dose-finding plays an important role in Phase II, designing a trial to find the correct dose is a clinical, as well as a statistical challenge. Another difficulty lies in the fact that Phase II studies are relatively short, so it is often not possible to measure long-term endpoints, which are of interest in the Phase III trials that follow. These long-term endpoints then have to be replaced by shorter-term endpoints (for example a biomarker, which can quickly be measured). Choosing the right endpoint and extrapolating the results correctly in this situation is a challenge with a clinical and a statistical component.

Phase III

The main purpose of Phase III trials (often called confirmatory trials) is to demonstrate efficacy of the treatment under study in a specific population that has a given disease [1]. The drug tested is compared to a control treatment (could be either placebo or standard-of-care) and patients are in most cases randomised. Such a trial should provide the definite proof of treatment superiority (or equivalence or non-inferiority, depending on the purpose of the trial). The sample sizes are often large, varying from a couple of hundred to even tens of thousands of patients. They also aim to identify long-term or rare side effects that have not been discovered in the previous phases. The aim of a Phase III trial is to provide sufficient evidence for the drug to be licensed by a regulatory agency for use in patients. However, nowadays only 25-30% of treatments gets approved and make it to patients [1].

The statistician is responsible for designing and analysing the trial. The design stage involves making decisions on questions like the sample size (the number of patients involved), the study design, the randomisation strategy (how the patients should be allocated to the different treatment options), the choice of endpoints (how the effect of the drug can be measured) and the statistical analysis methods.

After the trial is finished, the analysis of the trial data according to the SAP is performed. This involves not only the calculations that are necessary to decide if the new drug is better than, or as efficient, as the already approved drug or placebo treatment (primary endpoint), but also, for example, the validation of the data.

The adaptive design is considered as a highly innovative study design due to its flexibility and feasibility. In general, an adaptive (in design) clinical trial is a trial with a prescribed opportunity for modifications as the study progresses (e.g., in sample size, in dosage, in the hypothesis testing) based on analysis of interim data. Other innovations to the standard RCT (randomised controlled trial in which the subjects or groups of the subjects are randomly allocated) are the biomarker-marked designs (e.g., enrichment design) and multi-arm designs. The group sequential design also represents a ground-breaking design, as its main characteristic is the opportunity to stop the trial for futility or efficacy. However, the major scope for design innovations seems to be in Phase II.

The most important statistical methods seem to be the implementation of multiple testing (where a set of hypotheses are to be tested at the same time) and interim analyses (e.g., interpretation of early-stage results in the trial). In addition, survival analysis (e.g., analysing data where the time until the occurrence of an event such as death is observed) is considered as relevant, in particular related to the Kaplan-Meier method for estimating survival functions. Another relevant method is the regression analysis such as logistic regression (where the dependent variable representing the outcome of interest is categorical) and mixed model (where both random effects and fixed effects are included). Furthermore, methods for missing data as well as methods based on Bayesian approaches are also important for Phase III.

Stats terms in clinical trials - https://www.eupati.eu/clinical-development-and-trials/statistics-clinical-trials-key-concepts/

comparison structure of a trial and the

comparison structure describes the different ways that we compare

an experimental group to a control group in the trial.

The general types of comparison structures are

parallel, crossover and group allocation.

We'll start first with the parallel design.

This is the design that we usually think of when we think of a clinical trial.

In a parallel design, we are assigning patients and

administering treatment, so the experimental

and control groups in parallel.

In other words, we are assigning people to both

groups, over the same period of time, as opposed

to collecting data on the experimental groups

only, and comparing that data to historical

controls, or as opposed to assigning treatment

A and then assigning treatment B in series.

crossover design.

In a crossover design the unit that is randomized is the

order in which the treatments are received, instead of

whether or not the patient receives A or B.

So in a crossover design we randomize whether they receive

A first and then B or B and then A.

And so in this case, randomization promotes balance

between the treatment groups and timing of the exposure.

The defining feature of our crossover design is that we're testing each

treatment in all patients.

That means that each patient serves as his or her own control.

None of the treatments in the study can provide a permanent cure.

If a treatment has a permanent cure, then we can't cross

somebody over to the other treatment,

because they've already reached the outcome.

So the conditions for which we can use a

crossover design are only those that have a chronic

level of intensity for which the treatments

provide symptomatic relief but not any permanent cure.

In this first example, the investigators were comparing an evening

dose versus a morning dose of travoprost,

* [Cluster](https://en.wikipedia.org/wiki/Cluster_randomised_controlled_trial) – pre-existing groups of participants (e.g., villages, schools) are randomly selected to receive (or not receive) an intervention.

In the factorial design we are testing

two or sometimes more experimental interventions simultaneously.

We test the treatments simultaneously, either because

it's economical to test the two treatments simultaneously

or because the design can be used to

test for interaction between treatments A and B

Adaptive Designs:

And for fixed designs the design characteristics and

features such as the sample size, and the

hypotheses of interest, the outcomes, and the treatment

groups, those are all established before the trial starts.

That's not to say that there are never changes to a trial

with a fixed design once a trial starts.

In fact, changes do have to be made

because of information that you learn during the trial,

either from data within the trial or from relevant

data that is learned from outside of the trial.

However, some trials are designed to change

depending on what's observed in the trial.

And these are called adaptive designs.

There are many potential adaptations in adaptive designs.

Investigators can change randomization probabilities.

For instance, one might calculate the probability of success or improvement

in the outcome for the different treatment groups continually as participants move

through the trial.

And then this information can be used to adjust the

probability of being assigned to the different treatment groups so that

the next participant has a higher probability of being assigned

to the treatment group that's showing a higher probability of success.

Another adaptation is a change in the sample size based on the accruing data.

Group sequential methods, which are methods of stopping early due

to benefit or harm of a treatment, and also methods

for stopping early for futility have been around for some time.

And they are some of the most common and well understood design adaptations.

Also investigators might be unsure about

the best visit schedule for observing outcomes.

So they could specify a change if it

was discovered that the length of follow-up was

unnecessarily long or if they discovered that the

length of follow-up should be longer and the investigators

could also increase or decrease the number of interim follow-up visits.

You can also change the treatment groups during a trial.

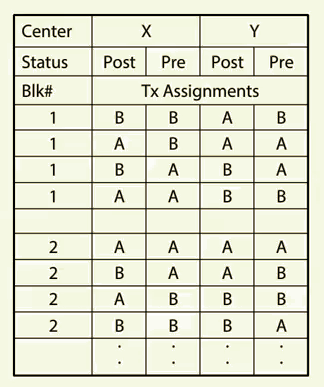
In order to do this, you need to specify rules under which a new

treatment can be added, or rules for when a current treatment can be dropped.

You might want to change the dose or duration of one or more of the treatments.

And you might decide to change the list of allowed or required concomitant meds.

Decide outcomes based on what the treatemtn might change – should be a proven outcome - should be something you can measure



"RCT" name only for trials that contain [control](https://en.wikipedia.org/wiki/Scientific_control) groups, in which groups receiving the experimental treatment are compared with control groups receiving no treatment (a [placebo-controlled study](https://en.wikipedia.org/wiki/Placebo-controlled_study)) or a previously tested treatment (a [positive-control study](https://en.wikipedia.org/wiki/Scientific_control#Positive)).

Randomization of patients: remove selection bias

* Simple - No idea what the next sample picked will be – but will get imbalance and reduced power
  + So you can create blocks of patients – and simple randomize that – you are assured that you have a balance
  + Stratify – based on clinic, gender, age, stage, etc – you want some amount of stratification so you can control for that – and then block!
  + Adaptive randomization – as the trial proceeds see if you want to only give one treatment or increase patients of a certain type

Masking:

* Treatment assignment not known even after the treatment – reporting is objective about how treatment works

The types of statistical methods used in RCTs depend on the characteristics of the data and include:

* For [dichotomous](https://en.wikipedia.org/wiki/Dichotomous) (binary) outcome data, [logistic regression](https://en.wikipedia.org/wiki/Logistic_regression) (e.g., to predict sustained virological response after receipt of [peginterferon alfa-2a](https://en.wikipedia.org/wiki/Peginterferon_alfa-2a) for [hepatitis C](https://en.wikipedia.org/wiki/Hepatitis_C)[[56]](https://en.wikipedia.org/wiki/Randomized_controlled_trial#cite_note-Manns-2001-56)) and other methods can be used.
* For continuous outcome data, [analysis of covariance](https://en.wikipedia.org/wiki/Analysis_of_covariance) (e.g., for changes in blood lipid levels after receipt of [atorvastatin](https://en.wikipedia.org/wiki/Atorvastatin) after [acute coronary syndrome](https://en.wikipedia.org/wiki/Acute_coronary_syndrome)[[57]](https://en.wikipedia.org/wiki/Randomized_controlled_trial#cite_note-Schwartz-2001-57)) tests the effects of predictor variables.
* For time-to-event outcome data that may be [censored](https://en.wikipedia.org/wiki/Censoring_(statistics)), [survival analysis](https://en.wikipedia.org/wiki/Survival_analysis) (e.g., [Kaplan–Meier estimators](https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator) and [Cox proportional hazards models](https://en.wikipedia.org/wiki/Cox_proportional_hazards_model) for time to [coronary heart disease](https://en.wikipedia.org/wiki/Coronary_heart_disease) after receipt of [hormone replacement therapy in menopause](https://en.wikipedia.org/wiki/Hormone_replacement_therapy_(menopause))[[58]](https://en.wikipedia.org/wiki/Randomized_controlled_trial#cite_note-Rossouw-2002-58)) is appropriate.

Drug and placebo patients are not reacting differently, drug A and drug B patients are not reacting differently- null hypothesis

Drug patients are reacting statistically significantly better than placebo – alternate hypothesis

Parametric tests – compare the means of two distribution – most commonly used

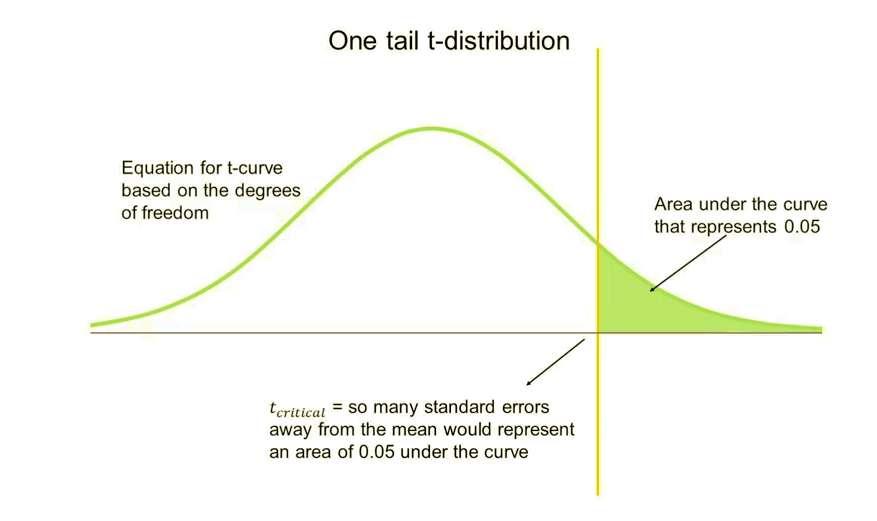
Students t-test

– the data has to be numerical, ratio, continuous, population should have normal dist

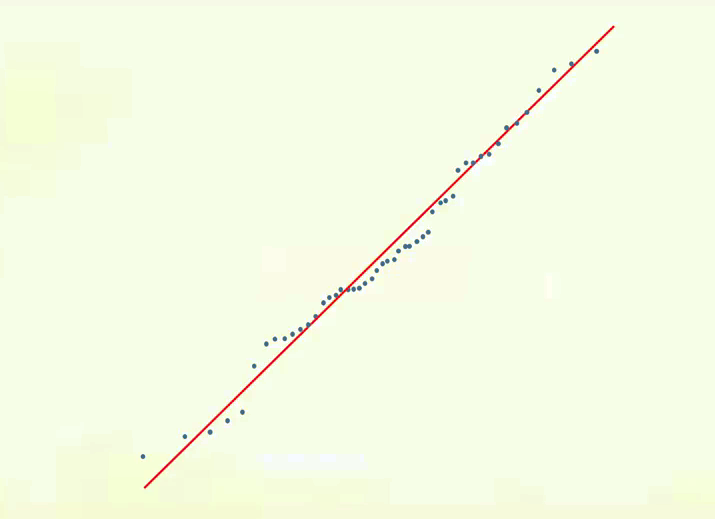
-Unpaired groups

-independent – cant be twins – cant be same patients who took one treatment once and then another treatment

– compare two distribution means



q-q plot –to check if normal distribution - each blue dot says how many points are less than this value – the red line should be like this for a normal distribution



Another classification of RCTs categorizes them as "superiority trials", "noninferiority trials", and "equivalence trials", which differ in methodology and reporting.[[36]](https://en.wikipedia.org/wiki/Randomized_controlled_trial#cite_note-Piaggio-2006-36) Most RCTs are superiority trials, in which one intervention is hypothesized to be superior to another in a [statistically significant](https://en.wikipedia.org/wiki/Statistical_significance) way.[[36]](https://en.wikipedia.org/wiki/Randomized_controlled_trial#cite_note-Piaggio-2006-36) Some RCTs are noninferiority trials "to determine whether a new treatment is no worse than a reference treatment."[[36]](https://en.wikipedia.org/wiki/Randomized_controlled_trial#cite_note-Piaggio-2006-36) Other RCTs are equivalence trials in which the hypothesis is that two interventions are indistinguishable from each other

In clinical trial.gov you have info about – disease name, treatment and how it is administered, patient info, study type (observational, interventional), number of pparticipants, randomization, masking, outcomes, what the patients preconditions are before accepted for trial, results, statistical anlaysis used

How would you reuse clinical data?

Two components to it:

1. What data is available and can we access it
2. Once we have the data what can we use it for – find targets, validate targets and also design trials

Data:

Clinical trials.gov, icgc – text mining

Pubmed, ehr - nlp

1

therapy or medication, Clinical info, biomarkers,

Study design, randomization, number of patients, Change precondiitons,

dosage, change times, analysis method

Lab results, documents from tests/reports/how that affected the patient, medication, vital signs, billing/provider info

ICGC data

Automated phenotype identification algorithms were developed using NLP techniques (to identify key findings, medication names, and family history), billing code queries, and structured data elements (such as laboratory results) to identify cases (n = 70–698) and controls

health care

provider notes, radiology reports, pathology reports, dis-

charge summaries, and operative reports

health care

provider notes, radiology reports, pathology reports, dis-

charge summaries, and operative reports

health care provider notes, radiology reports, pathology reports, dis-charge summaries, and operative reports - The variables included broad concept terms such asdisease diagnoses (RA, SLE, PsA, and JRA), medications(listed above, with the addition of adalimumab), labora-tory data (RF, anti-CCP, and the term “seropositive”), andradiology ﬁndings of erosions on radiographs. We used theHITEx system (19) to extract clinical information fromnarrative text. We extracted the variables mentioned abovefrom the narrative data and created coded NLP variablesfor the number of mentions per subject as well as dichot-omous variables for each disease diagnosis, medication,laboratory test result, and erosions on radiographs.

Toaccount for variability in language usage, a variety of spe-ciﬁc phrases can be deﬁned, which are then collapsed intoa single concept term for analyses. The clinicians on theteam developed lists of terms to be used for each NLPquery - For example, a patient wasﬂagged as being CCP positive by NLP if terms were foundin their records such as “anti-CCP⫹” and “CCP positiveRA.”

text mining and analysis:

* Theres 2 cases – structured and standardzed data – like clinical trials, icgc data portal – here you take unique words, lower case, stem, lemmatize, tokenize
* Unstructured and non standardized data – like EHR’s or scientific literature like pubmed databases – so like this therapy worked, patients condition improved, this outcome changed from this to this
* First thing you need to do in text mining is pre-process data – like stem workds, lemmatization (NLTK) – for unstructured data need to include gapped words, semantically similar (WordNet) or collocated words (NLTK bigram association measures)
* To analyze, you can try to get list of docs to review – using these words, topic modeling
* You can classify the docs, reports as this therapy worked, this did not work – svm (numeric features), naïve bayes
* To analyze, you can get features – which you can statistically analyze

<https://www.thieme-connect.com/products/ejournals/pdf/10.15265/IY-2017-007.pdf>

Analyses

Target identification and validation - what stats

* Identify biomarkers in cancer – this is critical both for target identification and for getting more success in clinical trials - Exciteing time cause there is so much data – that we can now access – and so many methods become available
* Do what I do – find individual biomarker associated with a disease - 21 SNPs that are known to be associated with these five diseases (atrial fibrillation, Crohn's disease, multiple sclerosis, rheumatoid arthritis, and T2D (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002823)>

Get large biomarker sets to predict outcome of a trial –also to get targets - immunotherapy – gut microbiome– new datasets, validation datasets

* Look at clinical trial data, icgc -Unsupervised - Do pca to see if a few features have the most variation – does pc1, pc2 explain the most variation?
* – can you regression, svm, to find combinations of biomarkers
* and new algorithms like apriori rf nn to find correlations on several levels – like individual genes, pathways

Clinical trials – what stats:

* linear regression and classification - to detect ADR signals from EHRs based on a drug/organ type
* From the patient data – come up with trajectories – given this symptom first, what is the next event and so on – until death
  + Disease trajectory -> Symptom – disease-death
  + Event trajectory -> Symptom – disease – admission – drug – surgery – death
* Tells you given this is the trajectory of patient – what is the probability of an event next – also can tell you time based on median time for other patients
* So maybe in cancer see how it spreads and figure out which combination of drugs might work
* Time of intervention is important
* This treatment – these side effects
* Classify saying for this cancer, this worked, did not work – this trial was successful and this wasnt – text mine clinical trial data - Study design, randomization, number of patients, patients precondiitons, dosage, analysis

identify patients for trials, market - Using a logistic regression algorithm operating on billing data, NLP-derived features, medication records, and laboratory data, Liao et al. developed an algorithm to accurately identify rheumatoid arthritis patients (https://onlinelibrary.wiley.com/doi/epdf/10.1002/acr.20184)

* Supervised – naïve bayes – say give this disease, how likely is this medicine to work compared to other medicines
* Unsupervised - Do pca to see if a few features have the most variation – does pc1, pc2 explain the most variation?
* Unsupervised - Hierarchical clustering to see if all the patients in a cluster have some feature
* Or cluster over features to see if tumor/normal fall in a particular feature
* logistic regression - to discover how the patient and the characteristics of support and intervention systems affect the improvement in urinary and bowel incontinence - to determine the association between nurse continuity and hospital-acquired pressure ulcers
* In orthogonal regression, the relationship between the dependent and the independent variable(s) is the one that minimizes the orthogonal distances from the observed values of the dependent variable and the corresponding values on the fitting line. Sun and colleagues used orthogonal regression to identify risk factors related to an adverse condition
* Boolean logic extracts data using queries made by Boolean combinations of a set of conditions. Boolean logic was applied in many studies, i.e., [157] and [158], ranging from the analysis of EHRs for the evaluation of the effectiveness of triage models used in mass casualty research to the identification of emergent endotracheal intubation in ICU patients.
* Fuzzy logic is used to solve problems where it is more convenient to consider the concept of ‘partial truth’: a variable might be partially true or partially false. An example is given in [159] where EHRs are analyzed to detect potential ADR signals
* The Apriori algorithm is the most widely known association rule algorithm using an iterative approach to find the most frequent associations between two or more items and gives a measure of the frequency with which that particular association has been found. The algorithm has been applied in [168] to discover associations between diagnoses of different sub-groups of patients. Association rule mining has been applied in [169] to identify the associations between combination of diagnoses, demographics, and lab results to predict high risk of diabetes. In [170] association rule was applied to discover medical correlations, characterize data trends, and perform predictive analysis on data trends and medical correlations.
* Classification And Regression Tree (CART) analysis, a particular type of decision tree, has been applied to detect ADRs
* k-NN is used in [173] for retrieving patients with similar characteristics by analyzing EHRs
* Fuzzy neural networks are the combination of neural networks and fuzzy logic. Skevofilakas and colleagues used fuzzy neural networks to predict the risk of Type I Diabetes Mellitus patients to develop diabetic retinopathy
* Support Vector Machines (SVM) aim at assigning a new observation into one of two possible categories. It was applied in combination with Bayesian networks and k-NN in [175] to predict pancreatic cancer
* fuzzy-clustering is used for the identification of rare-cases in post-operative pain management
* hierarchical clustering has been applied in [177] to identify periodic/seasonal patterns in incidence of diseases
* Probabilistic graphical models, such as Bayesian networks, are a widely used class of structured prediction models. Graphic models describe the underlying relations between the variables with a graph: the links between the different variables represent the conditional dependencies between the variables. Bayesian networks together with k-NN and SVM were used in [175] to predict pancreatic cancer by using knowledge-base from PubMed research papers and experimental observations derived from EHRs. Graphic modeling is found also in [182] to identify which user accesses to EHR data deviate from the accesses found during typical patient care
* Topic modeling relies on statistical models for extracting the “topics” that occur in a set of documents. One of the models used in topic modeling is the Latent Dirichlet Allocation (LDA) where the statistical information is assumed to have a Dirichlet distribution. LDA was used in [183] for EHR-driven phenotyping
* PhWas – figuring out if a patient has more than 1 disease
* The PheWAS algorithm, then calculates case and control genotype distributions and calculates the χ2distribution, associated allelic P-value and allelic odds ratio (OR). For those χ2 distributions in which observed cell counts fell below five, Fisher's exact test was used to calculate the P-value

Reuse of clinical data:

The benefits of reusing clinical information have been well documented in the clinical literature for decades [[5-7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287069/#ref5)]. Cohort analysis has been used to determine risk for readmission to the hospital within 30 days from discharge [[8](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287069/#ref8)]; predict death and length of stay based upon abnormal laboratory values [[9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287069/#ref9)]; describe populations of patients [[10](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287069/#ref10), [11](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287069/#ref11)]; assist in infection control [[12-18](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287069/#ref12)]; and discover pharmaco-epidemiological relationships

To improve healthcare management and quality, clinical data has already been reused to measure and improve quality [23,24], predict patients length of stay, discharge, readmission, and death [25-28], and improve infection control [29-31]. Data has also been reused for early detection of diseases, pharmacovigilance, and post-market and public health surveillance [32]. In clinical research, data has been reused to accelerate and increase patient recruitment in trials [33], enable in-silico hypothesis testing [34], and enable faster and cheaper access to a richer variety of clinical information for various types of clinical research applications such as comparative effectiveness research and patient phenotype combination with genomic data. As discussed by Coorevits and colleagues, clinical data reuse “will optimize research and development platforms, processes, and timelines”, will generate “high-quality clinical evidence faster through better protocol feasibility assessment, improved patient identification and recruitment, and more efficient clinical study conduct, including for reporting serious adverse events”, “will maximize the value to customers and diversify revenue streams” of research organizations, and enable the participation of clinical investigators and physicians in a larger number of clinical trials

Clinical registries are usually well structured; meaning the sites contributing data use standardized forms and controlled vocabularies.

aggregations of clinical data from EHRs such as the General Practice Research Database (GPRD) [[11](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287069/#ref11)] in the United Kingdom are less well structured, but are close to real-time.

Moreover, the data density across patients is not regular; data are missing. Clinicians choose when to order tests so not every similar patient will have a particular test at a particular time. Clinicians also introduce bias when they select treatments without providing the reason for selection.

When data are aggregated from more than one clinical setting, the meaning of data may not be consistent. For instance, if a cardiologist records chest pain on a problem list he might mean something different from a gastroenterologist who has recorded the same problem for a different patient.

Unstructured data case:

<https://www.i2b2.org/NLP/DataSets/Main.php>

https://www.sciencedirect.com/science/article/pii/S1532046415000891?via%3Dihub

With the advice of practicing medical doctors and researchers, we developed an annotated corpus that answers the question “For each record in each patient’s EMR, which [cancer](https://www.sciencedirect.com/topics/medicine-and-dentistry/cardiovascular-disease) risk factors (family history, smoker etc) were present before, during, and after the record’s creation date?” – what were the biomarkers like before, during and after – this is a group of patients

– these aer the biomarkers – what were they treated with – what worked

- dosages, modes, frequencies, durations, reasons

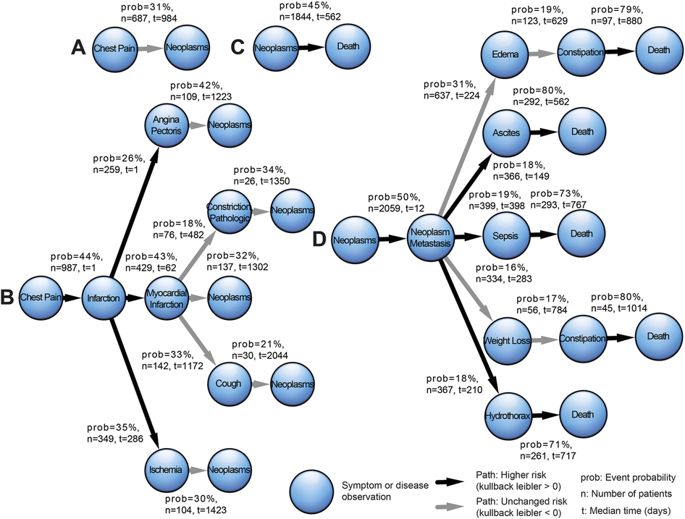
- drug-side effects

- drug induced reactions

The 2010 i2b2/VA shared task [[6]](https://www.sciencedirect.com/science/article/pii/S1532046415000891?via%3Dihub" \l "b0030) had three tracks: (1) concept extraction, where systems had to identify patient medical problems, treatments, and tests; (2) assertion classification, where identified concepts from the previous track were categorized as being present, absent, possible, conditional, hypothetical, etc.; and (3) relation classification, where relationships between the concepts were categorized into types. For example, medical problems could relate to tests in several ways including “test reveals medical problem”, “test conducted to investigate medical problem”, or “in the same sentence but the relationship is other/undefined”. 871 medical records were annotated for the 2010 shared task.

* Identify the risk factors - Keep track of time
* A light annotation implements the following principles: it employs expert annotators; it generates few, document-level, evidenced-based annotations – like only a positive correlation with disease – and quantitative measurement, not high bmi
* Check your annotations using training data and test data
* You can get both descriptive (just a summary of observations) and inferential (inferred to a larger sample) stats
* Not all the patients that respond to therapy will have the same marker or level of marker – so give percentage of patients that respond – you will need to use statistics to determine if that is significant – what is the control?

https://www.nature.com/articles/srep46226



Data scientist GMD:

* They’ll ask generally did she do good, creative, broad stats
* Gen stats questions
* Addressing these 4 projects and late stage clinical trial ideas
* values

Electronic Heath

Imaging

Bio marker integration

Marketing to diff patients

Gen late stage trials – it’s the usual stats – plus how to determine patient size

* assay development, measurement method comparison, biomarker cut-off determination and validation - biomarkers (e.g. cut-off decision, bridging studies)
* evaluation of new predictive and prognostic biomarkers

We are looking for a Research Segment Strategies Analyst - Field for our Research Business Division, which includes customer institutions from universities, government, hospitals, pharma and biotech industries.

Your remit will be to segment our customer base and identify high potential sub-segments, based on insights from various data points, such as data analytics, market trends, customer journey and geo-mapping, and help develop targeted approaches to capture business opportunities.

* Drive novel ways to approach customer segmentation
* Identify high growth and underserved sub-segment areas
* Leverage both internal and external data to arrive at fact-supported segmentation recommendations
* Explore  global and regional market dynamics and variance to prioritize target sub-segmentsConduct customer interviews to obtain insights
* Understand top subsegments’ customer journey, their pain points and needs, and make recommendations for customer programs development

Electronic Heath  - see notes

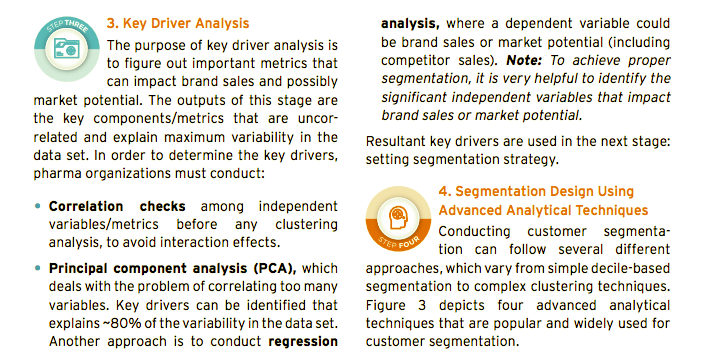
Bio marker integration – merck

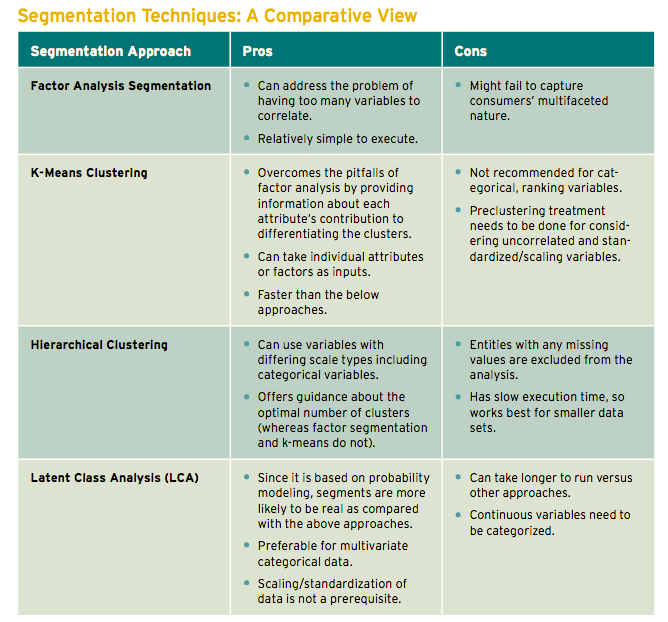
And apriori algorithm?

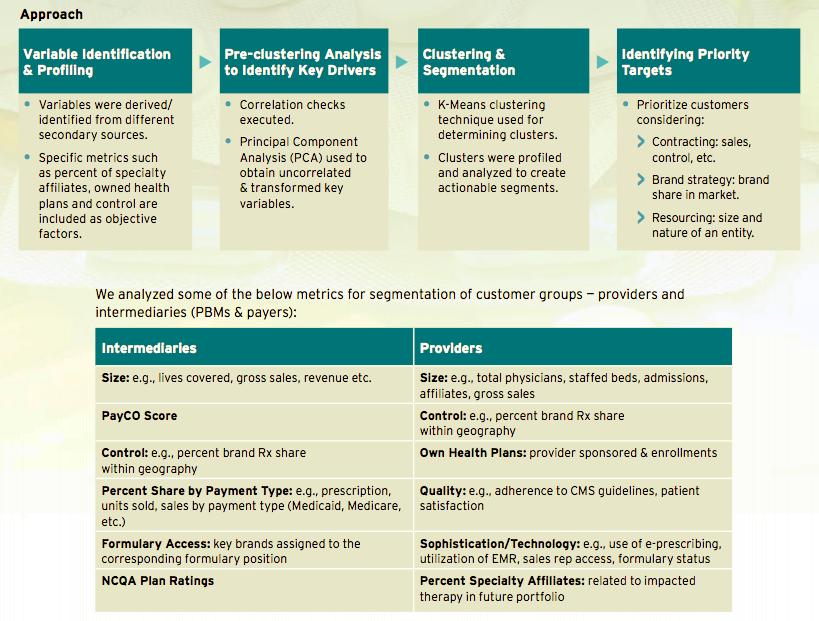
Marketing to diff patients

* Known to work on patients – but there are several competitor options available
* Figure out who your customers are – payers, providers, direct to patient
* For each segment figure out how much of this drug AZ is selling based on things like – size, number of patients affected, number of drugs selling – can get this from EHR and you can do text mining
* These are all the features - across the board say these are correlated with people buying AZ – in places where we miss patients – yea we are missing these – could be type of advertising, payer commitment
* pca – include patients who are buying AZ and other brands – pca should pick up the factors that are most variable between the groups
* regression – say advertising spend across different advertising channels is different – as you increase spend on a certain type of advertising – say direct sales teams – do you see AZ being used more – R2
* classification – you don’t need a model to classify patients here cause you don’t need to classify a new patient – you need to know the connection between chossing AZ – and if the connection is significant
* clustering – for patients who do buy AZ and other brands – are the patients clustered in a certain geography – then you would reach out to those physicians more – or advertise to those areas more
* use tests after the regression, classification – to see if significant differences between groups - for number of patients who buy AZ and other brands – contingency table with payers or marketing channel on the y axis – fishers, chi squared
* Why if we capture most of the market and why if less – advertising spend - like is the core patient group a certain age, race - like a certain distribution/marketing channel - like a certain brand historically - Can get this info from EHR

1. Finding potential customer segments – how much of that segment we tapped







1. Existing payers, physicians, direct to consumer
   1. EHR– cluster by geography – to see if more patients in that region based on doc recommendation can take that drug
   2. group incentives
   3. age, ethinicity, sex etc.
   4. based on browsing history – if a patient searched for liver cancer
   5. based on what a person buys
   6. Maybe one part of the high prescribing group is very convinced by journal articles; maybe another group is more convinced by KOLs.
   7. I have seen approaches based on usage segmentation, behavioral segmentation (both prescribing and social), attitudinal segmentation (e.g., innovation), psychographic segmentation (lifestyle, choices, personality differences), demographic segmentation (age, gender), geographic segmentation (regional, rural, urban), and more
   8. Channels to market on – journals, speaker programs, direct sales teams
2. Also new patient populations:
3. Common biomarkers – Across diseases - <https://hmpi.org/wp-content/uploads/2017/02/HMPI-Trusheim-Berndt-ONC-segmentation.pdf>

(Gleevec: Novartis) has been proven beneficial in Philadelphia chromosome positive CML and acute lymphatic leukemia, Kit (CD117) positive gastrointestinal cancer and mastocytosis with specific c-Kit mutations and PDGFR (platelet derived growth factor receptor) gene rearrangements, but has not been successful in biomarker sub-types of CRC, NSCLC, liver, renal, ovarian, thyroid and head and neck cancers.

Trastuzumab (Herceptin: Roche/Genentech), initially targeted HER2 3+ overexpressing breast cancer. In the 12 years since its approval, its use has only been extended to one other organ type: metastatic gastric or GEJ (Gastroesophageal junction) cancer

In addition to organ – interactions within the organ should be considered

Focus on job search – will take 3-4 months to finish – plenty of time to finish running and on the side – write after for a couple months and take a month vaca!

Once you have an offer – expedite and finish interviews - talk to a ton of people to see which job to take based on criteria below – and also the team lead

Criteria for me:

* Big company
* Can move to US
* Position match my experience
* skills and area onco – can learn new needed things
* make 120k so youre not upset
* you need to use your current experience but also learn about other biomarkers and clinical analysis? – TALK TO PEOPLE

after you get job – while at duke – learn up any additional skills you need – finish work here

take a 1 month vacation

new place – keep looking for opportunities in the US – in about 6 months see if it’s a good time to tell them that you’ll move in future