A data pipeline approach to orphan disease insights

- KNIME 2018 FALL SUMMIT -

Sebastien Lefebvre

Senior Director – Data Analytics & Decision Support

Alexion pharmaceuticals Inc

ALEXION

Key Themes in Artificial Intelligence in Rare Disease Diagnosis

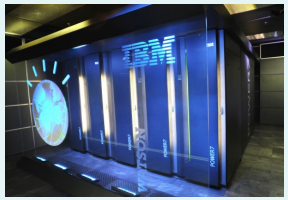
Impact

2018

A new world record of 19.5 hours is set for the fastest genetic diagnosis by Rady Children's Institute for Genomic Medicine.

http://www.frontlinegenomics.com/review/21973/65-years-of-dna/

Computational Learning



- **Computational hypothesis generation, data interpretation, decision support, and acceleration of human insight to enable rare disease diagnosis**

Manage bias

Information Fusion

- **Semantically integrate and navigate complex, heterogeneous, local and distributed data**

Manage ambiguity

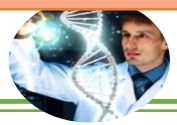
"Big Data"

- **Capture, process, filter, and manage a global and growing avalanche of internal and external scientific and clinical data**

Manage variety

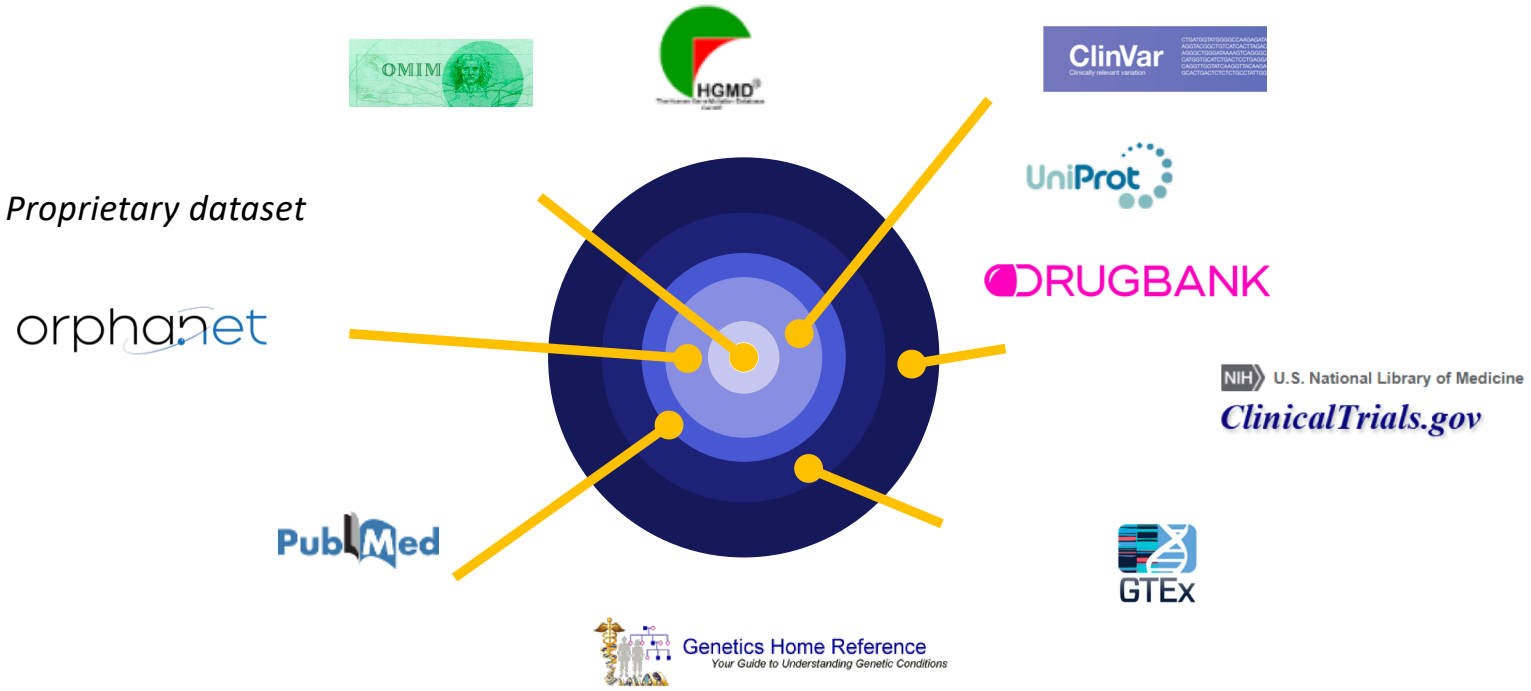
Source: John Reynders VP data sciences at Alexion

AI Decision Support



The Alexion Insight (AI) Engine

MAPPING AND INTEROGATING THE UNIVERSE OF RARE DISEASES

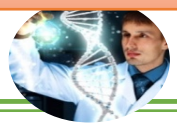


Each **link** represents a set of evidences that are extracted using a complex data pipeline

Open for Innovation KNIME

amazon web services

AI Decision Support



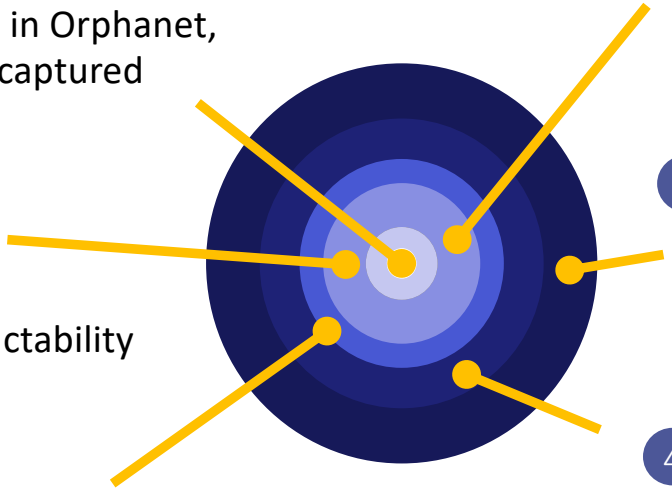
The Alexion Insight (AI) Engine

MAPPING AND INTEROGATING THE UNIVERSE OF RARE DISEASES

1 **All Rare Diseases**
- There are approximately 9,500 rare diseases reported in Orphanet, OMIM, and GHR – all captured our data graph

2 **Incidence/Prevalence**
- Mine registry and genetic information
- Insight into disease tractability
- Patient segmentation

3 **Disease Onset**
- Clarify window of opportunity and identify patient populations



6 **Genetic Architecture**
- Detailed mapping of all known genetic mutations in the target disease with proprietary bioinformatics

5 **Competitive Intensity**
- Estimate competitive landscape via our novel scoring scheme using drugs, clinical trials, medical specialties…

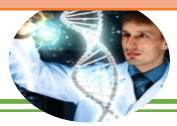
4 **Severity and unmet needs**
- Find the most severe diseases via phenotype ontology and age of death
- Standard of care

Each **link** represents a set of evidences that are extracted using a complex data pipeline

Open for Innovation KNIME

amazon web services

AI Decision Support



The Alexion Insight (AI) Engine

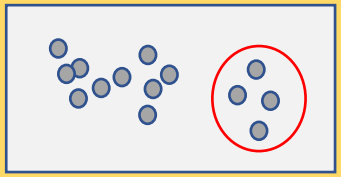
ANSWERING KEY RARE DISEASE QUESTIONS

HGMD · OMIM · orphanet · ClinVar · PubMed · GTEx · DRUGBANK · NIH U.S. National Library of Medicine ClinicalTrials.gov · UniProt · Genetics Home Reference Your Guide to Understanding Genetic Conditions

Proprietary datasets

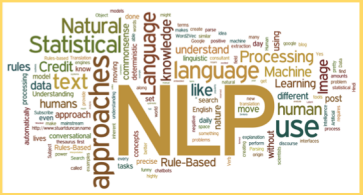
XML, JSON, CSV, RestAPI, SQL…

Enforce appropriate comparisons*

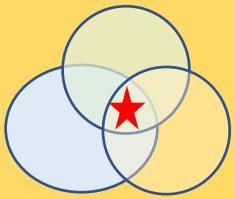


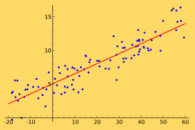
Inspect & evaluate alternative explanations*

Document the source and characteristic of the data*



Demonstrate mechanism of cause & effect and express them quantitatively*





Recognize the inherent multivariate nature of the problem*

Conditions of interest

Clinical asset deal flow

Orphan disease diagnostics

KOL mapping

Patient population size and segmentation

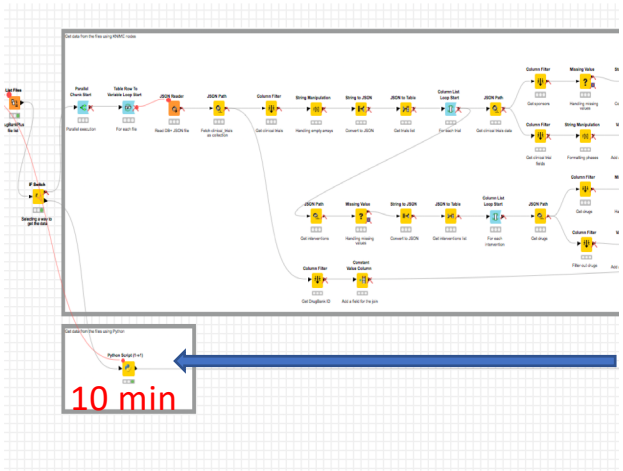
Each **question** is answered with a dedicated data analytics pipeline highlighting key insights

Open for Innovation KNIME · amazon web services

*Edward Tufte on visual design of a data analysis

JSON parsing example

Thousands of complex JSON files




```
import pandas
import json
import glob

def get_clinical_trials(filename):
    with open(filename) as drugbank_file:
        drugbank_data = json.load(drugbank_file)
        drugbank_id = drugbank_data.get('drugbank_id')
        trials = drugbank_data.get('clinical_trials')
        clinical_trials = []
        if trials:
            for trial in trials:
                sponsors = trial.get('sponsors')
                if not sponsors:
                    add_clinical_trial(clinical_trials, drugbank_id, trial)
                else:
                    for sponsor in sponsors:
                        add_clinical_trial(clinical_trials, drugbank_id, trial, sponsor)
        return clinical_trials

def get_all_clinical_trials():
    clinical_trials = []
    for filename in glob.iglob(flow_variables['drugbankplus_path'] + '/**/DB*.json', recursive=True):
        clinical_trials.extend(get_clinical_trials(filename))
    return clinical_trials

def get_drug_value(value_name, interventions):
    return set([drug_name for intervention_drug_names in [[drug.get(value_name) for drug in intervention.get('drugs')]
                                                          for intervention in interventions]
                for drug_name in intervention_drug_names])

def add_clinical_trial(clinical_trials, drugbank_id, trial, sponsor = None):
    conditions = trial.get('conditions')
    condition_titles = '|'.join(set([condition.get('title') for condition in conditions])) if conditions else None
    condition_synonyms = '|'.join(filter(None, [('|'.join(condition.get('synonyms')) if condition.get('synonyms') else None)
                                                for condition in conditions])) if conditions else None
    interventions = trial.get('interventions')
    intervention_kinds = '|'.join(set([intervention.get('kind') for intervention in interventions])) if interventions else None
    intervention_titles = '|'.join(set([intervention.get('title') for intervention in interventions])) if interventions else None
    drug_names = '|'.join(get_drug_value('name', interventions)) if interventions else None
    drug_drugbank_ids = '|'.join(get_drug_value('drugbank_id', interventions)) if interventions else None
    clinical_trials.append({'drugbank_id': drugbank_id, 'identifier': trial.get('identifier'), 'status': trial.get('status'),
                            'title': trial.get('title'), 'official_title': trial.get('official_title'), 'purpose': trial.get('purpose'),
                            'phases': '|'.join(map(str, trial.get('phase'))) if trial.get('phase') else None,
                            'start_date': trial.get('start_date'), 'end_date': trial.get('end_date'),
                            'brief_summary': trial.get('brief_summary'),
                            'condition_titles': condition_titles, 'condition_synonyms': condition_synonyms,
                            'intervention_kinds': intervention_kinds, 'intervention_titles': intervention_titles,
                            'drug_names': drug_names, 'drug_drugbank_ids': drug_drugbank_ids,
                            'sponsor_title': sponsor.get('title') if sponsor else None,
                            'sponsor_agency_class': sponsor.get('agency_class') if sponsor else None,
                            'lead_sponsor': sponsor.get('lead_sponsor') if sponsor else None})

output_table = pandas.DataFrame(get_all_clinical_trials())
```

Postgres table

Next: Big Data extension → Spark JSON?

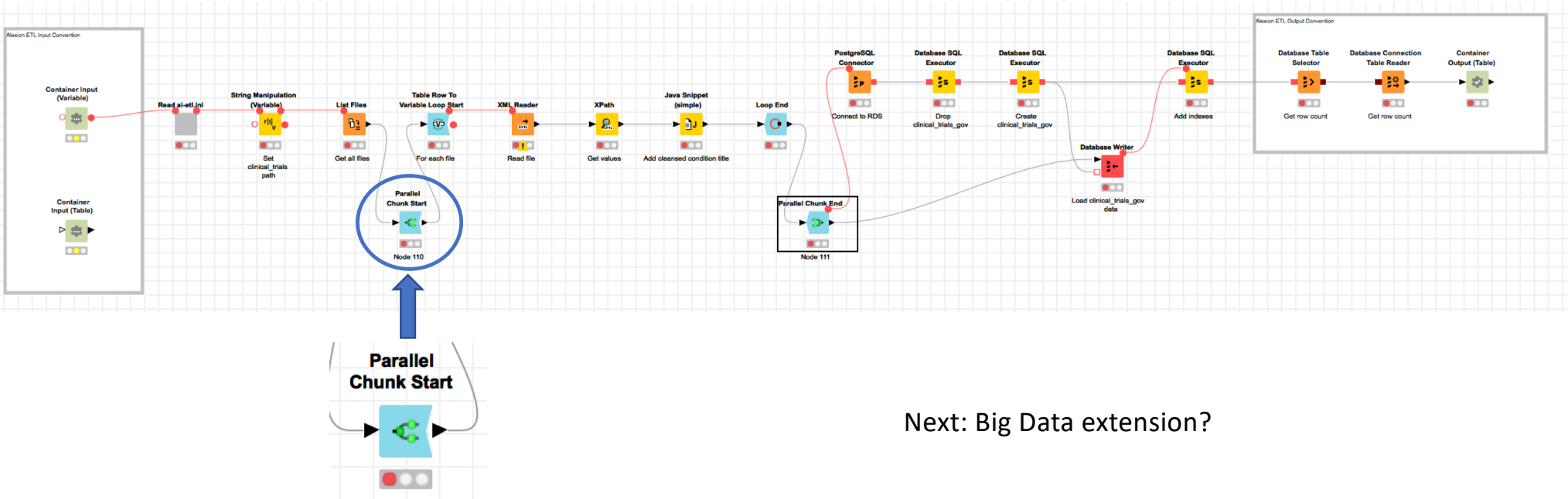
Clinical trials example

286,199 XML files

Originally 35+ hours

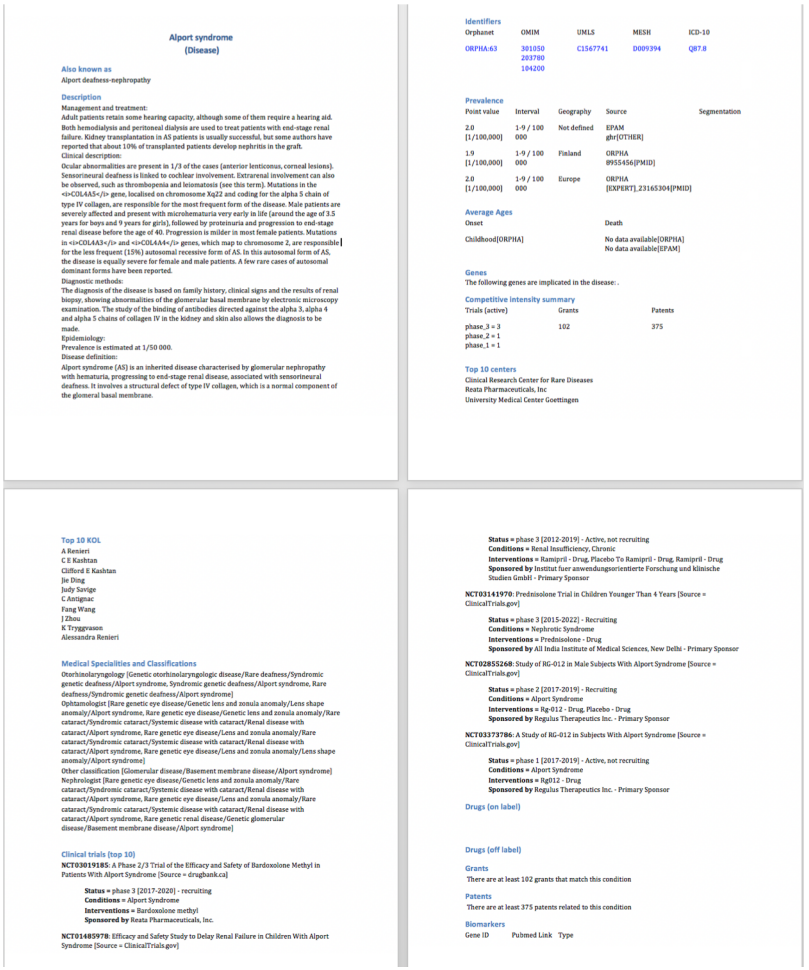
With *parallel chunk* node 2.5 hours

Postgres table



Next: Big Data extension?

Disease card example



Alport syndrome
(Disease)

Also known as
Alport deafness-nephropathy

Description
Management and treatment:
Adult patients retain some hearing capacity, although some of them require a hearing aid.
Both hemodialysis and peritoneal dialysis are used to treat patients with end-stage renal failure. Kidney transplantation in AS patients is usually successful, but some authors have reported that about 10% of transplanted patients develop nephritis in the graft.
Clinical description:
Ocular abnormalities are present in 1/3 of the cases (anterior lenticonus, corneal lesions). Sensorineural deafness is linked to cochlear involvement. Extrarenal involvement can also be observed, such as thrombopenia and leiomatosis (see this term). Mutations in the COL4A5 gene, localised on chromosome Xq22 and coding for the alpha 5 chain of type IV collagen, are responsible for the most frequent form of the disease. Male patients are severely affected and present with microhematuria very early in life (around the age of 3.5 years for boys and 9 years for girls), followed by proteinuria and progression to end-stage renal disease before the age of 40. Progression is milder in most female patients. Mutations in COL4A3 and COL4A4 genes, which map to chromosome 2, are responsible for the less frequent (15%) autosomal recessive form of AS. In this autosomal form of AS, the disease is equally severe for female and male patients. A few rare cases of autosomal dominant forms have been reported.
Diagnostic methods:
The diagnosis of the disease is based on family history, clinical signs and the results of renal biopsy, showing abnormalities of the glomerular basal membrane by electronic microscopy examination. The study of the binding of antibodies directed against the alpha 3, alpha 4 and alpha 5 chains of collagen IV in the kidney and skin also allows the diagnosis to be made.
Epidemiology:
Prevalence is estimated at 1/50 000.
Disease definition:
Alport syndrome (AS) is an inherited disease characterised by glomerular nephropathy with hematuria, progressing to end-stage renal disease, associated with sensorineural deafness. It involves a structural defect of type IV collagen, which is a normal component of the glomeral basal membrane.

Identifiers

Orphanet	OMIM	UMLS	MESH	ICD-10
ORPHA:63	301050 203780 104200	C1567741	D009394	Q87.8

Prevalence

Point value	Interval	Geography	Source	Segmentation
2.0 [1/100,000]	1-9 / 100 000	Not defined	EPAM ghr[OTHER]	
1.9 [1/100,000]	1-9 / 100 000	Finland	ORPHA 8955456[PMID]	
2.0 [1/100,000]	1-9 / 100 000	Europe	ORPHA [EXPERT]_23165304[PMID]	

Average Ages

Onset	Death
Childhood[ORPHA]	No data available[ORPHA] No data available[EPAM]

Genes
The following genes are implicated in the disease: .

Competitive intensity summary

Trials (active)	Grants	Patents
phase_3 = 3 phase_2 = 1 phase_1 = 1	102	375

Top 10 centers
Clinical Research Center for Rare Diseases
Reata Pharmaceuticals, Inc
University Medical Center Goettingen

Top 10 KOL
A Renieri
C E Kashtan
Clifford E Kashtan
Jie Ding
Judy Savige
C Antignac
Fang Wang
J Zhou
K Tryggvason
Alessandra Renieri

Medical Specialities and Classifications
Otorhinolaryngology [Genetic otorhinolaryngologic disease/Rare deafness/Syndromic genetic deafness/Alport syndrome, Syndromic genetic deafness/Alport syndrome, Rare deafness/Syndromic genetic deafness/Alport syndrome]
Ophtamologist [Rare genetic eye disease/Genetic lens and zonula anomaly/Lens shape anomaly/Alport syndrome, Rare genetic eye disease/Genetic lens and zonula anomaly/Rare cataract/Syndromic cataract/Systemic disease with cataract/Renal disease with cataract/Alport syndrome, Rare genetic eye disease/Lens and zonula anomaly/Rare cataract/Syndromic cataract/Systemic disease with cataract/Renal disease with cataract/Alport syndrome, Rare genetic eye disease/Lens and zonula anomaly/Lens shape anomaly/Alport syndrome]
Other classification [Glomerular disease/Basement membrane disease/Alport syndrome]
Nephrologist [Rare genetic eye disease/Genetic lens and zonula anomaly/Rare cataract/Syndromic cataract/Systemic disease with cataract/Renal disease with cataract/Alport syndrome, Rare genetic eye disease/Lens and zonula anomaly/Rare cataract/Syndromic cataract/Systemic disease with cataract/Renal disease with cataract/Alport syndrome, Rare genetic renal disease/Genetic glomerular disease/Basement membrane disease/Alport syndrome]

Clinical trials (top 10)
NCT03019185: A Phase 2/3 Trial of the Efficacy and Safety of Bardoxolone Methyl in Patients With Alport Syndrome [Source = drugbank.ca]

> **Status =** phase 3 [2017-2020] - recruiting
> **Conditions =** Alport Syndrome
> **Interventions =** Bardoxolone methyl
> **Sponsored by** Reata Pharmaceuticals, Inc.

NCT01485978: Efficacy and Safety Study to Delay Renal Failure in Children With Alport Syndrome [Source = ClinicalTrials.gov]

> **Status =** phase 3 [2012-2019] - Active, not recruiting
> **Conditions =** Renal Insufficiency, Chronic
> **Interventions =** Ramipril - Drug, Placebo To Ramipril - Drug, Ramipril - Drug
> **Sponsored by** Institut fuer anwendungsorientierte Forschung und klinische Studien GmbH - Primary Sponsor

NCT03141970: Prednisolone Trial in Children Younger Than 4 Years [Source = ClinicalTrials.gov]

> **Status =** phase 3 [2015-2022] - Recruiting
> **Conditions =** Nephrotic Syndrome
> **Interventions =** Prednisolone - Drug
> **Sponsored by** All India Institute of Medical Sciences, New Delhi - Primary Sponsor

NCT02855268: Study of RG-012 in Male Subjects With Alport Syndrome [Source = ClinicalTrials.gov]

> **Status =** phase 2 [2017-2019] - Recruiting
> **Conditions =** Alport Syndrome
> **Interventions =** Rg-012 - Drug, Placebo - Drug
> **Sponsored by** Regulus Therapeutics Inc. - Primary Sponsor

NCT03373786: A Study of RG-012 in Subjects With Alport Syndrome [Source = ClinicalTrials.gov]

> **Status =** phase 1 [2017-2019] - Active, not recruiting
> **Conditions =** Alport Syndrome
> **Interventions =** Rg012 - Drug
> **Sponsored by** Regulus Therapeutics Inc. - Primary Sponsor

Drugs (on label)

Drugs (off label)

Grants
There are at least 102 grants that match this condition

Patents
There are at least 375 patents related to this condition

Biomarkers
Gene ID Pubmed Link Type

AI Decision Support

Insights must be FAIR*

Findable, Accessible, Interoperable, Reusable

Package your insight with a version:

- **Description of the question/problem** with assumptions/inclusions (e.g. presentation, publication)
- A readable **data analysis story** supporting the insight (e.g. KNIME workflow, *jupyter* notebook)
- **Data/Model snapshot** used for the analysis (e.g. python *pickle* file, flat file export, Model file)
- Any supporting **dataset** (e.g. gene-disease mapping, knowledgebase timestamp, training set)
- **Code** used to run the analysis (e.g. label in *GitHub*/*Artifactory*, python virtualenv)
- **Environment** that ran the code (e.g. AWS AMI, python virtualenv)
- Use/re-use **templates/packages** (e.g. KNIME metanode, *Artifactory ,* python virtualenv)

You will want to know how you came up with a given insight

Your Commercial, BD or Corp Strategy partner will ask!…trust me

Place the "Description of the analysis" inside one of your enterprise document repository so your work can be found by enterprise tools like *Search* and *ExpertFinder.*

* https://www.go-fair.org/fair-principles/

ALEXION