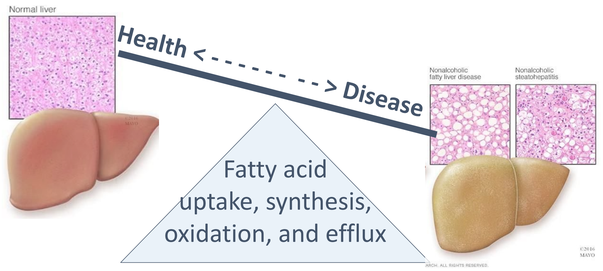
Development of a curated hepatic steatosis (HS) database to enable Quantitative Structure-Activity Relationship modeling

Nyssa N. Tucker[1], Vinicius M. Alves[2], Eugene Muratov[2], Alexander Tropsha[2]

[1] Biological and Biomedical Sciences Program, UNC Chapel Hill , [2] Molecular Modelling Lab, UNC Chapel Hill

Introduction

- Hepatic steatosis, also known as non-alcoholic fatty liver disease, is characterized by abnormal fat accumulation in the liver.
- Disease spectrum impacts **one in three adults** and **one in ten children** in the US.
- It is a multifactorial disease and common causes include environment, diet, behavior, and genetics.



Four apical key events serve as the fulcrum potentiating additional disease outcomes of HS.

- HS can evolve to further adverse outcomes, including fibrosis, cirrhosis, and cancer.

This study aims to collect, curate, and integrate the largest chemogenomics HS dataset to enable the accurate identification of novel potential HS causing agents.

Materials and methods

We performed extensive literature and web search and compiled data from:

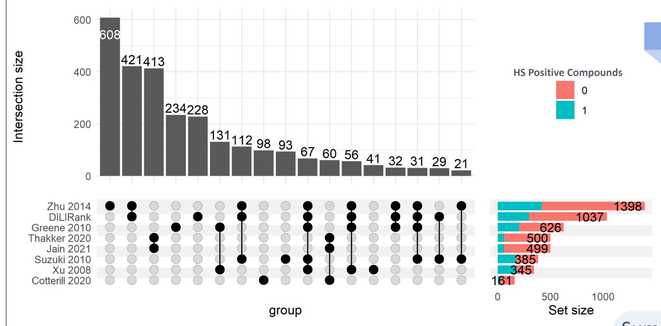
- Publications identified in PubMed
- Supplementary materials
- Publicly accessible electronic databases
- Private contributions

Data integration, curation, analysis, and visualization was executed in R.

Results and Discussion

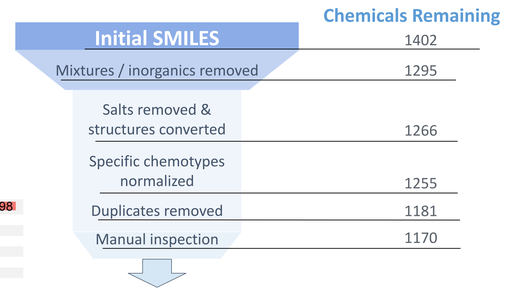
Data overview and curation

Overlapping analysis



Chemical data overlap between top source datasets, visualized by count overlapping and entire set size color coded by HS positivity (0 = negative, 1 = positive).

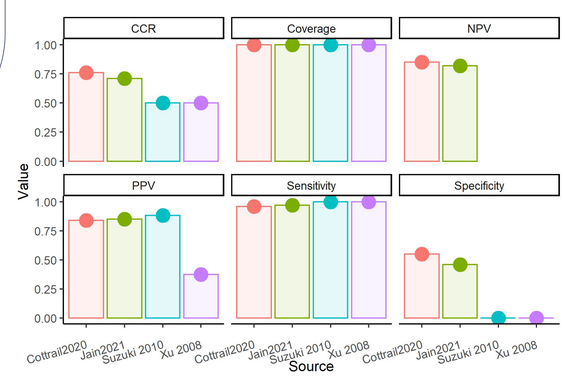
Key steps of chemical curation



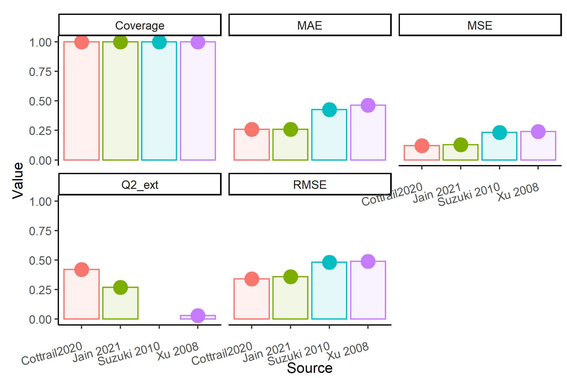
Summary of the chemical curation workflow, modified from [Fourches 2016]. Initial curation executed using subset of identified sources: Jain et al. 2021, Cotterill et al. 2020, Xu et al. 2008, and Suzuki et al. 2011.

Data analysis

Classification results



Regression results



Conclusions

Succeeded in developing the largest publicly available HS database. Using this database, we have developed QSAR HS models. Preliminary results suggest applying alternative regression strategy, as well as revisiting upstream workflow.

Future Directions

Data analysis

- Hierarchical cluster analysis to identify scaffolds associated with HS.
- Evaluation of overlap and concordance on conserved chemical subsets between data sources.

Cheminformatics analysis and modeling

- Cluster chemicals and analyze SAR to identify motifs related to HS.
- Identify representative chemicals causing additional adverse outcomes.
- QSAR Modeling
  - Virtual screening of chemicals of interest
  - Model interpretation to identify statistically validated chemical moieties associated with HS.

Experimental validation

- Validate computational models using *in vitro* assays with a collaborator at EPA.

Key References

- Angrish et al. 2016 10.1093/toxsci/kfw018
- Fourches et al. 2016 10.1021/acs.jcim.6b00129
- Tropsha 2010 10.1002/minf.201000061

virtual poster

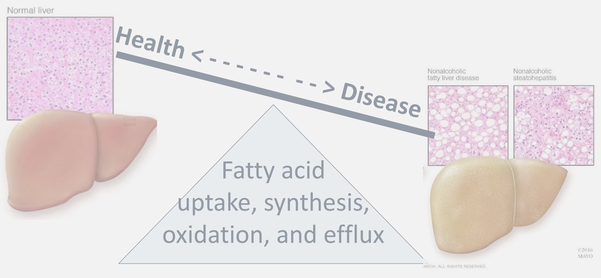
Development of a curated hepatic steatosis (HS) database to enable Quantitative Structure-Activity Relationship modeling

Nyssa N. Tucker[1], Vinicius M. Alves[2], Eugene Muratov[2], Alexander Tropsha[2]

[1] Biological and Biomedical Sciences Program, UNC Chapel Hill , [2] Molecular Modelling Lab, UNC Chapel Hill

Introduction

- Hepatic steatosis, also known as non-alcoholic fatty liver disease, is characterized by abnormal fat accumulation in the liver.
- It affects **one in three adults** and **one in ten children** in the United States.
- It is a multifactorial disease and common causes include environment, diet, behavior, and genetics.



Four apical key events serve as the fulcrum potentiating additional disease outcomes of HS.

- HS can evolve to further adverse outcomes, including fibrosis, cirrhosis, and cancer.

This study aims to collect, curate, and integrate the largest chemogenomics data on HS to enable the development of QSAR models that can help identify novel potential HS-causing agents.

Materials and methods

We performed extensive literature and web search and compiled data from:

- Publications identified in PubMed
- Supplementary materials
- Publicly accessible electronic databases
- Private contributions

Data integration, curation, analysis, and visualization was executed in R.

Results and Discussion

Data overview

Data extracted from 14 published HS-QSAR studies (subset below)

Author/Year	Organism	Type	Total compounds	Positive	Negative
Zhu 2014	human	in vivo	2029	1315	714
Thakker 2020	human	mix	1279	768	511
Jain 2021	rodent	in vivo	1041	120	921
Greene 2010	mix	in vivo	1266	640	626
Xu 2017	human	in vitro	333	155	178
Shin 2020 -d	human	in vivo	318	192	126
Shin 2020 -m	human	in vivo	254	152	102
Cotterill 2020	rodent	in vivo	164	82	82
Cotterill 2020	human	in vivo	42	21	21

Existing literature identifying common sources of chemical descriptors and classifiers.

Assay types

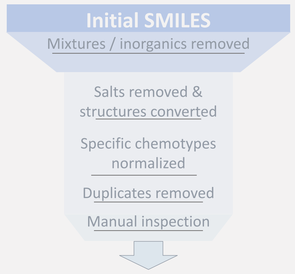
14 total publications

Clinical	In Vitro	Mix
Human: 8	Human: 4	Human: 1
Murine: 2	Murine: 0	Murine: 0

Accumulated HS data split by species, assay type.

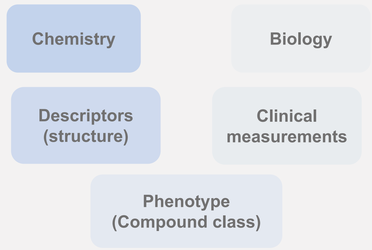
Data curation and analysis

Key steps of chemical curation



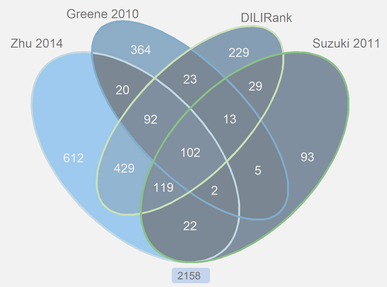
Summary of the chemical curation workflow. Modified from [Fourches 2016].

Descriptive features



Spectrum of source information characterizing chemicals bridging chemical and biological areas.

Overlapping analysis



Chemical data overlap between top source datasets.

Conclusions

Chemically induced HS poses a major health risk. The development of a curated HS database will enable the development of predictive QSAR models that will support early stage identification of chemicals inducing HS.

Future Directions

Data analysis

- Hierarchical cluster analysis to identify scaffolds associated with HS.
- Evaluation of overlap and concordance on conserved chemical subsets between data sources.

Cheminformatics analysis and modeling

- Cluster chemicals and analyze SAR to identify motifs related to HS.
- Identify representative chemicals causing additional adverse outcomes.
- QSAR Modeling
  - Calculate chemical descriptors
  - Development of predictive QSAR models using modern artificial intelligence algorithms
  - Virtual screening of chemicals of interest
  - Model interpretation to identify statistically validated structural alerts associated with HS.

Experimental validation

- Validate computational models using *in vitro* assays with a collaborator at EPA.

Key References

- Angrish et al. 2016 10.1093/toxsci/kfw018
- Fourches et al. 2016 10.1021/acs.jcim.6b00129
- Tropsha 2010 10.1002/minf.201000061

virtual poster