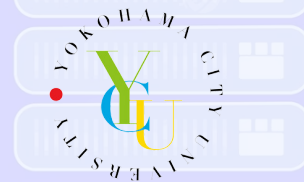


# 大学院講義「先端医科学研究概論」

臨床ビッグデータ解析のための公開 Web ツールの紹介

臨床ビッグデータ解析のための一般公開されているウェブツールの紹介

佐倉 絵里 前園 絵里  
バイオインフォマティクス 助教



2023-10-25

# 私のプロフィール

## プロとしての経験

**バイオインフォマティクス** 助教 2023年4月～現在  
横浜市立大学先端医療研究センター

**バイオインフォマティクス** 研究者 2020-2023 年 3 月  
Craif Inc.、日本、分析チーム

## 教育

筑波大学 2015-2020  
博士課程修了 ヒューマンバイオロジー  
大学院統合・グローバル専攻博士課程修了

## スキル

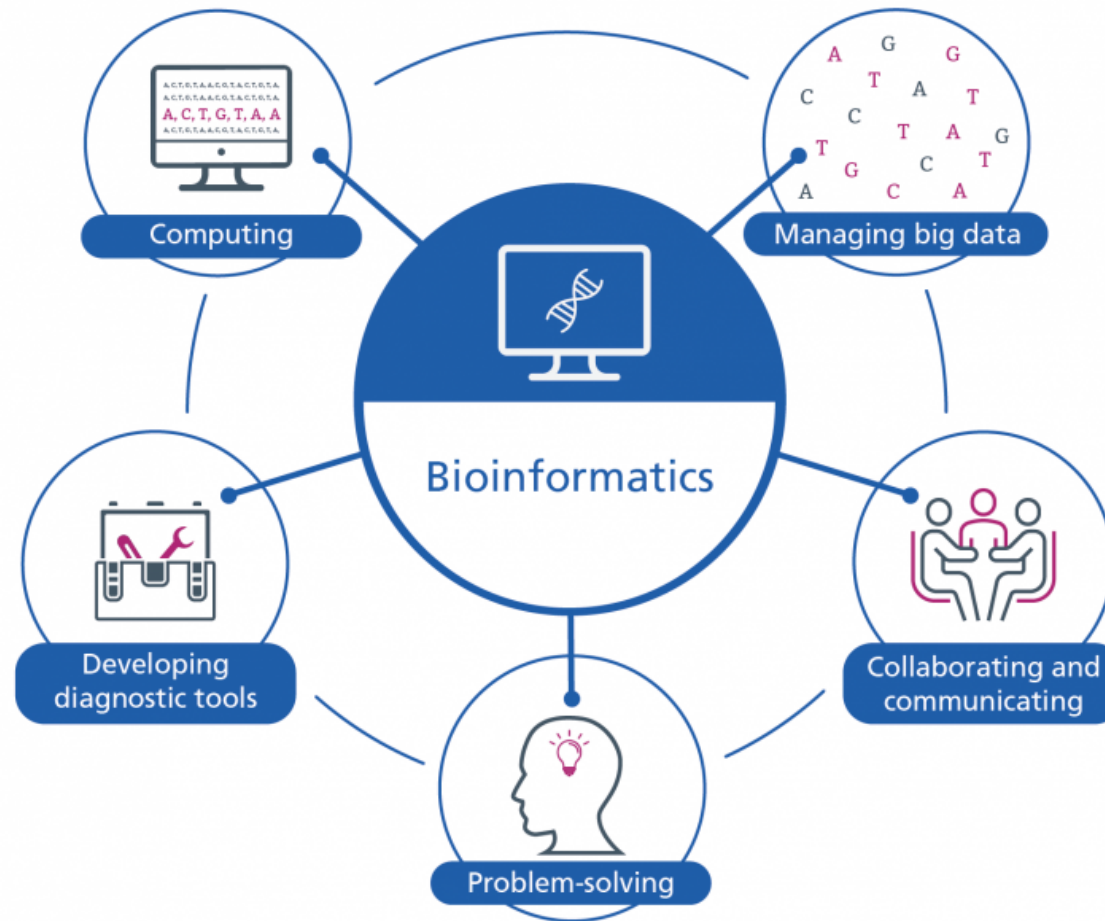
研究設計・統計解析の計画と実施・臨床データ管理・R/Shiny/Bioconductor・  
Python/Streamlit/Django・JavaScript(Google apps script)・Gitを利用したバージョン管理・機械学習



さくら 前園  
絵理博士



# バイオインフォマティクスとは何ですか？



# 私のプロフィール

## プロとしての経験

バイオインフォマティクス助教 2023年4月～現在  
横浜市立大学 先端医療研究センター

バイオインフォマティクス研究者 2020-2023 年 3 月  
Craif Inc.、日本、分析チーム

## 教育

筑波大学 2015-2020  
博士課程修了 ヒューマンバイオロジー  
大学院統合・グローバル専攻博士課程修了

## スキル

研究設計・統計解析の計画と実施・臨床データ管理・R/Shiny/Bioconductor・  
Python/Streamlit/Django・JavaScript(Google apps script)・Gitを利用したバージョン管理・機械学習



さくら 前園  
絵理博士







## 人々が天寿を全うする社会の実現

Craifは、尿を利用したあらゆる疾患の早期発見と治療最適化を目指しています。  
エクソソーム回収技術のパイオニアとして、精確で痛みのない疾患の早期発見方法を確立し  
誰もが生涯にわたって健康でいられる社会を実現します。

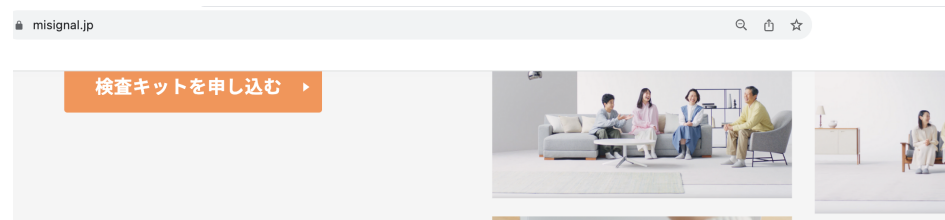
# 過去の研究 | がんリスク評価キット: Craif miSignal®

## 7つのがんの種類

- 食道
- 肺
- 胸
- 胃
- 膵臓
- 結腸直腸
- 卵巣

## 責任

- 臨床データ管理
- miRNA発現データの前処理
- バイオマーカー発見分析
- 予測アルゴリズム開発



### 特徴 01

#### 最大7種類のがんリスクをまとめて検査できます

わずかな尿を使って、特に発症数・死亡数が多いがん最大7種類のリスクをまとめて検査。

すい臓がんや卵巣がんなど、がん検診(対策型検診)でカバーされていないがん種のリスクの検査もできます。

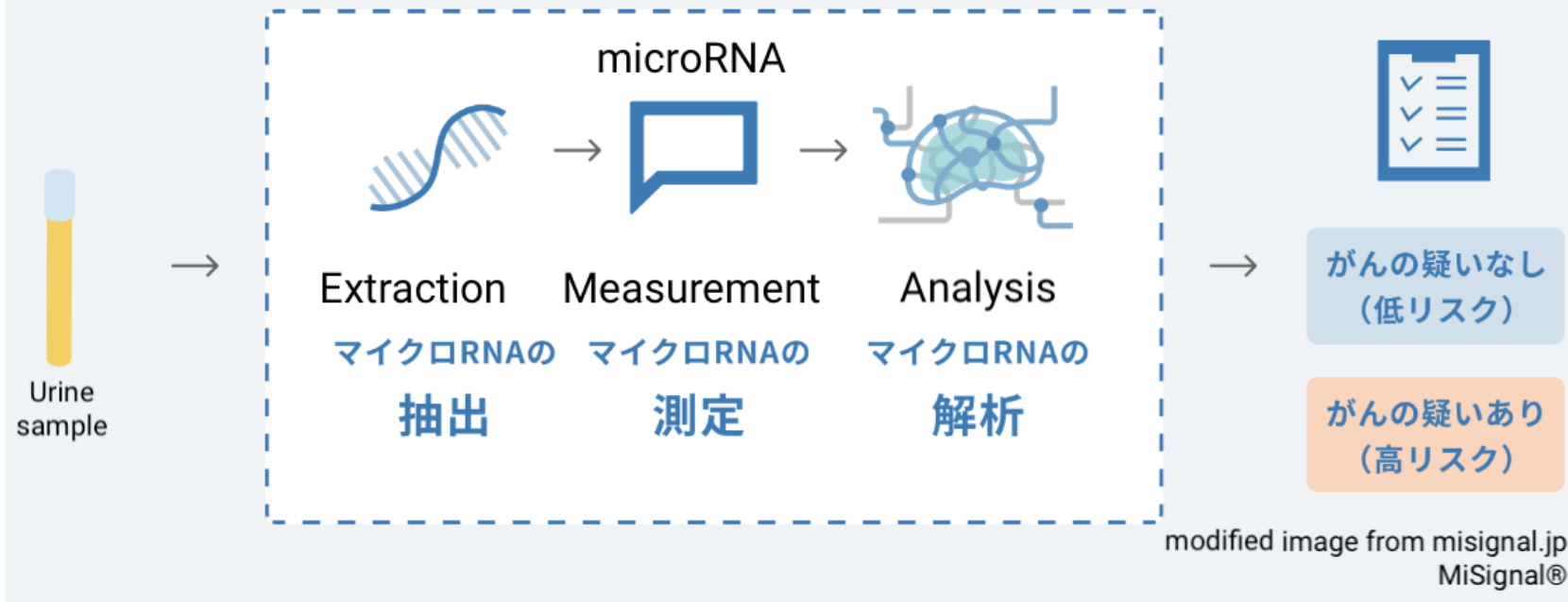


### 特徴 02

#### がんの種類ごとにリスクがわかる だから万が一の時も安心

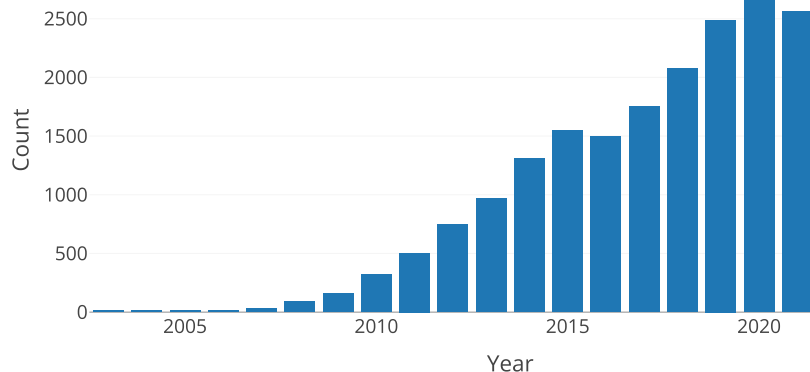


## マイシグナルのがんリスク検査の プロセスと3つの技術



# 過去の研究 | がんバイオマーカーとしてのマイクロRNA

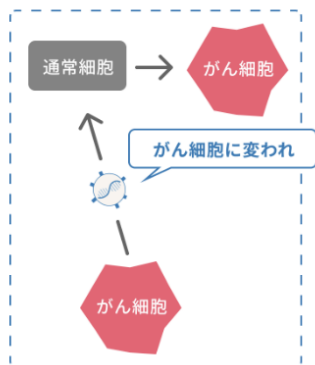
22265 Pubmed results for  
'miRNAs as cancer biomarkers'



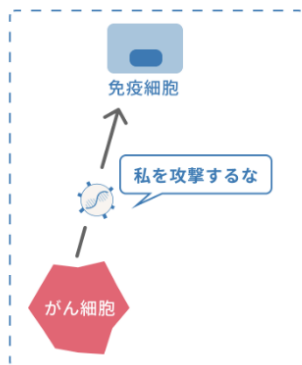
## どうやって？

- 遺伝子発現の調節への関与
- 診断、分類、予後、治療反応など、**がんのさまざまな側面との関連性**
- その**多用途性と非侵襲的検出により**、腫瘍学における貴重なツールとなっています。

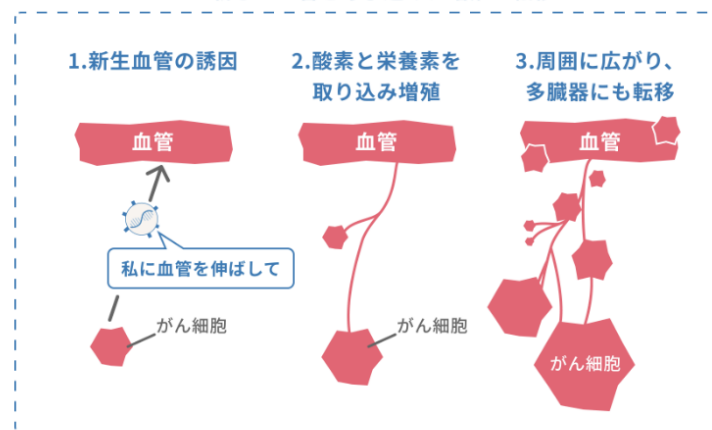
### A. 性質の変化（細胞のがん化）



### B. がんを破壊しようとする免疫細胞の抑制



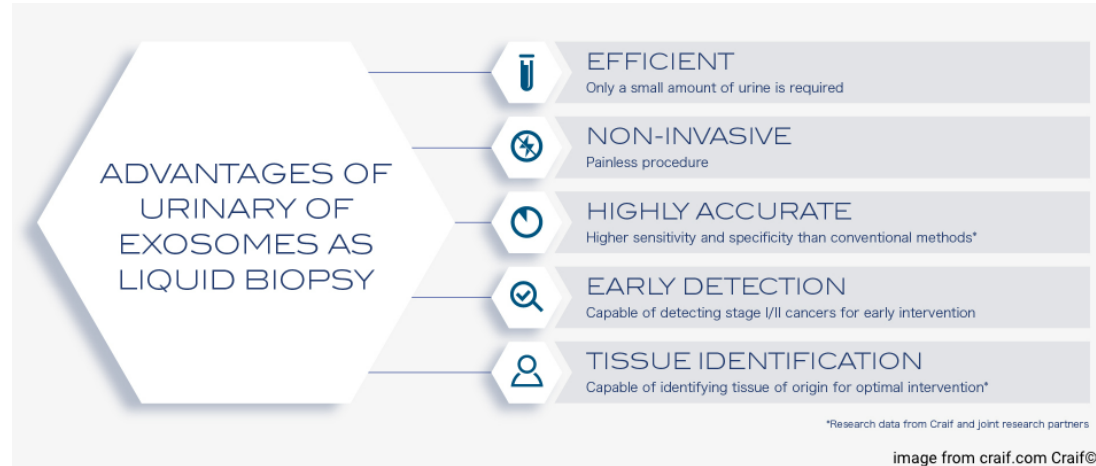
### C. 新しい血管を呼び込み、増殖・転移



images from misignal.jp MiSignal®

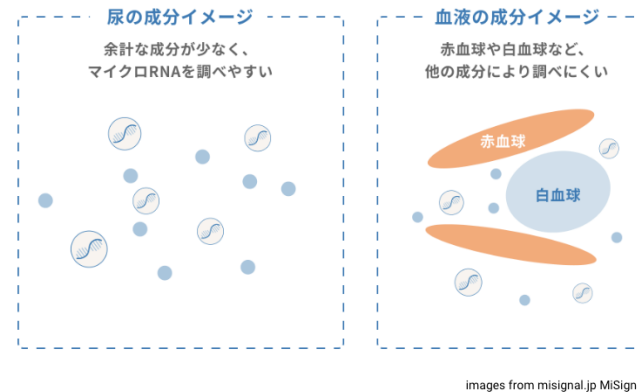


# 過去の研究 | なぜ尿なのか？



## なぜ血ではないのですか？

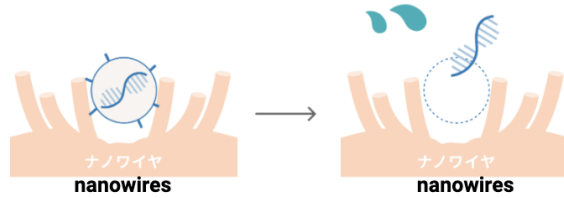
- 注射針が必要 (非侵襲的)
- より低い miRNA が検出されました (~600)。タンパク質などの不純物は検出を妨げる可能性があります (vs. >1300; 腎臓による濾過により汚染物質はほとんどありません)



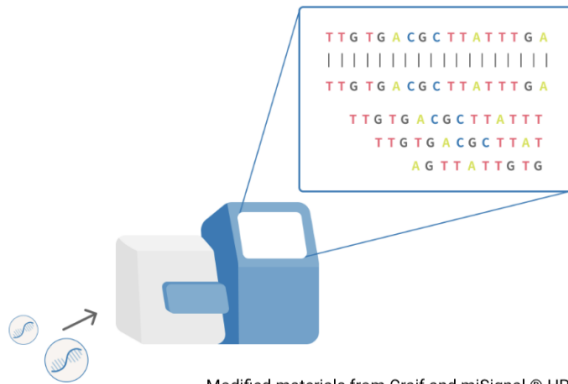


# 過去の研究 | miSignal® の開発におけるバイオインフォマティクスの役割

## 1. RNA Extraction

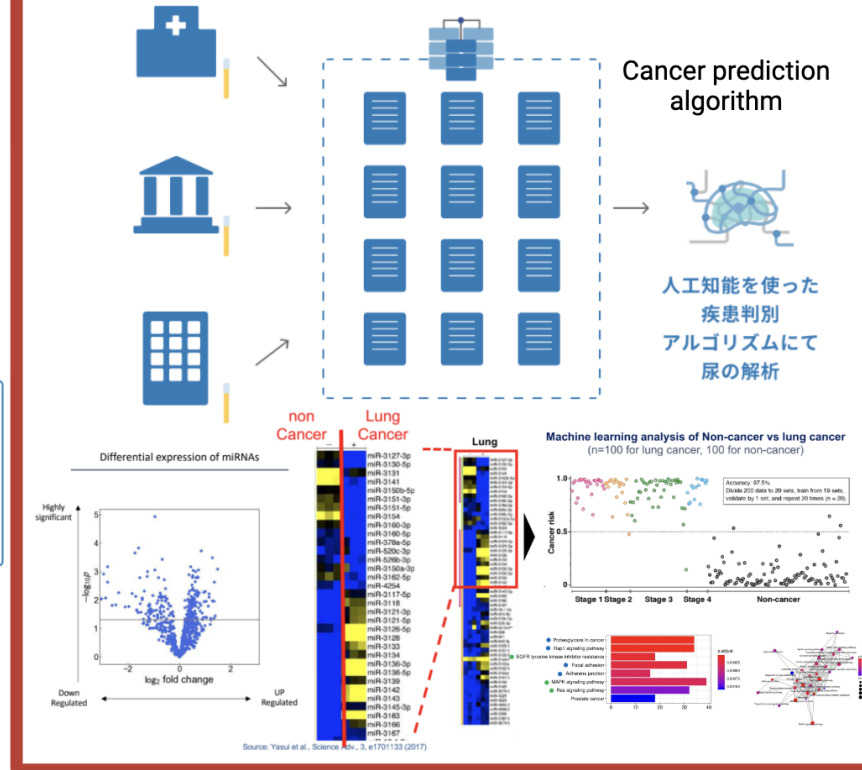


## 2. Measurement



## 3. Analysis (Prediction)

Multiple Hospitals/  
Institutions      A database of >10,000  
urine samples





# 私のプロフィール

## プロとしての経験

バイオインフォマティクス 助教 2023年4月～現在  
横浜市立大学 先端医療研究センター

バイオインフォマティクス研究者 2020-2023 年 3 月  
Craif Inc.、日本、分析チーム

## 教育

筑波大学 2015-2020  
博士課程修了 ヒューマンバイオロジー  
大学院統合・グローバル専攻博士課程修了

## スキル

研究設計・統計解析の計画と実施・臨床データ管理・R/Shiny/Bioconductor・  
Python/Streamlit/Django・JavaScript(Google apps script)・Gitを利用したバージョン管理・機械学習



さくら 前園  
絵理博士



# バイオインフォマティクス研究室

<https://www-user.yokohama-cu.ac.jp/~bioinfo/>

## YCU Bioinformatics Laboratory

🏠 Home   👤 Members   🧪 Research   📰 News  
🏫 Education   ❤️ Recruitment

[Yokohama City University](#) (YCU) is an important and innovative Research & Educational Institution in Japan. [Advanced Medical Research Center](#) (AMRC/SENTANKEN) was established in 2006 at YCU Fukuura Medical Campus to further promote medically-relevant research through various industry and academic collaborations. **Bioinformatics Laboratory** was formed in 2009 as an integral part of AMRC.

## Our Mission

- 🧬 [biologically-relevant research](#)
- 📖 [bioinformatics education](#)
- 🤝 [collaborations & internationalization](#)



# バイオインフォマティクス教育ポータル

<https://edu.med.yokohama-cu.ac.jp/>

The screenshot shows a dark-themed website with a sidebar on the left and a main content area on the right. The sidebar contains navigation links: Home, Our Courses, R Resources, Bioinformatics Tools, and Related Links. Below these are the YCU logo, Bioinformatics Laboratory name, and social media links for Github and E-Mail. The main content area features a welcome message, a 'Bioinformatics Courses' section with a description of online courses, an 'R Resources' section about R programming for biological data, and a 'Bioinformatics Tools' section with a 'Coming soon!' message.

Education Portal

Bioinformatics Laboratory

Home

Our Courses

R Resources

Bioinformatics Tools

Related Links

YOKOHAMA UNIVERSITY  
YCU

Bioinformatics Laboratory

Github E-Mail

Laboratory Home

バイオインフォマティクス  
ポータル

Welcome to YCU Bioinformatics Laboratory Education Portal!

This Education Portal is brought to you by [YCU Bioinformatics Laboratory](#). We hope you find many useful resources here and Join our Courses. Happy browsing!

Bioinformatics Courses

All our courses are online. Practical courses are also recorded and available on-demand. Open to YCU Students (free), Researchers, Collaborators and to External Participants. Please check Bioinformatics Courses to learn more about Current Courses and Apply!

R Resources

Efficient analysis of biological data in R requires basic coding skills and using RStudio. It is also recommended to take advantage of extensive R ecosystem offering over 20,000 packages to help us with data analysis. Please see the R Resources section to learn more!

Bioinformatics Tools

Coming soon! Visit the Portal for more details.

# 講義の内容

パート1 臨床ビッグデータ分析の概要

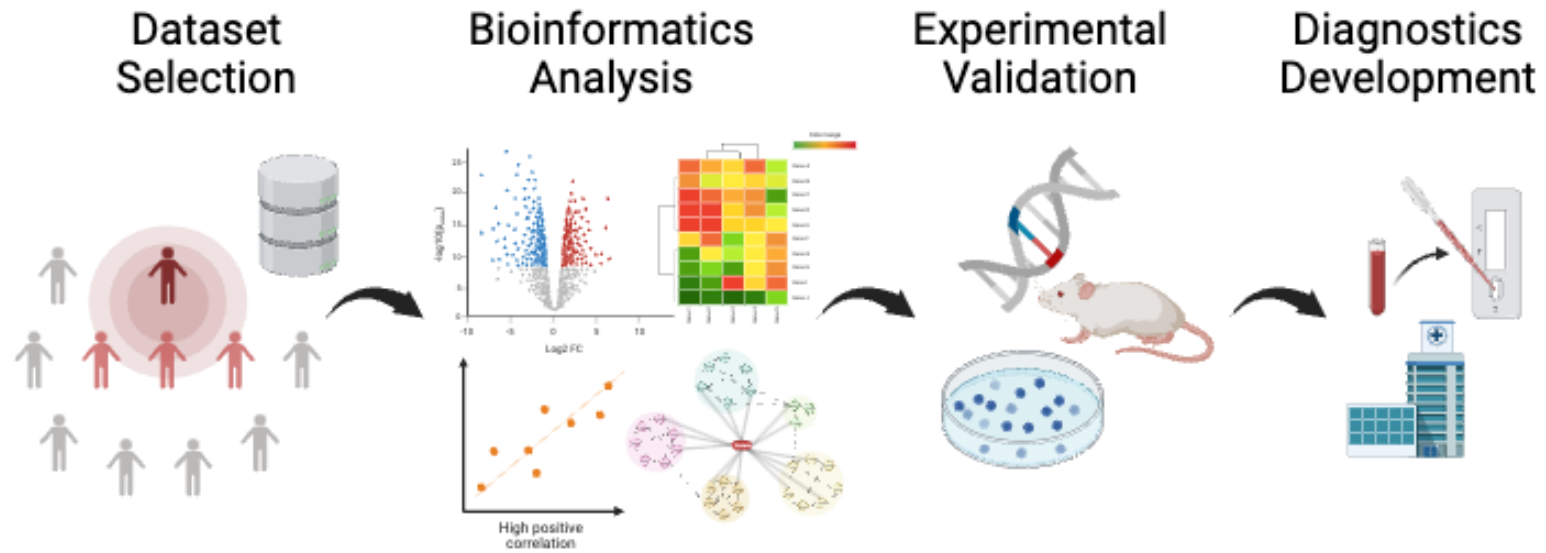
パート2 バイオインフォマティクス Web ツール

第3部 臨床ビッグデータ解析の課題と将来

# パート 1: 臨床ビッグデータ分析の概要

# What is Clinical Big Data Analysis?

the process of extracting valuable insights from vast and diverse datasets related to healthcare and medicine



# Importance of Clinical Big Data Analysis

The **complete sequencing of the human genome** has helped to unlock the genetic contribution for many diseases  
Its applications include the following:

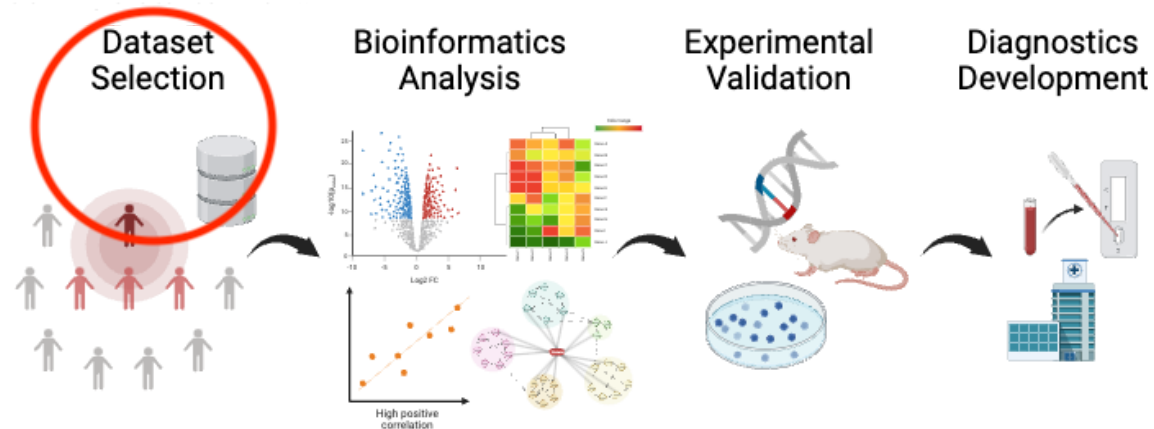
- **drug discovery:** Drug target identification and drug candidate screening can be accelerated, and safer/more effective drugs can be developed based on **molecular modelling and simulation**

- **personalized medicine:** A patient's **genetic profile** can assist the doctor to predict susceptibility to certain diseases, provide proper medication, and with the proper dose to reduce side-effects

- **gene therapy:** Identifying the **best gene target site** for each individual by taking their genetic profile into consideration can reduce the risk of unintended side effects

- **preventive medicine:** Genomics, proteomics, and metabolomics data are analyzed for possible **disease biomarkers** to develop screening tests that identify the disease at an early stage

# It always starts with...



## Types of data involved

- **Patient Data:** Demographic information, lifestyle factors, and health-related behaviors
- **Electronic Health Records (EHRs):** Comprehensive patient records, including medical history, treatment, and lab results
- **Genomics Data:** Information about an individual's genetic makeup, including DNA sequences and variations



# Available Big Data (databases) and how to access them

## Clinical

- store and manage data related to patients' medical history, diagnoses, treatments, and outcomes
  - Surveillance, Epidemiology, and End Results (SEER)
  - National Health and Nutrition Examination Survey (NHANES)
  - The Cancer Proteome Atlas (TCPA)
  - The Cancer Genome Atlas (TCGA)
  - Medical Information Mart for Intensive Care (MIMIC)

## Genomic

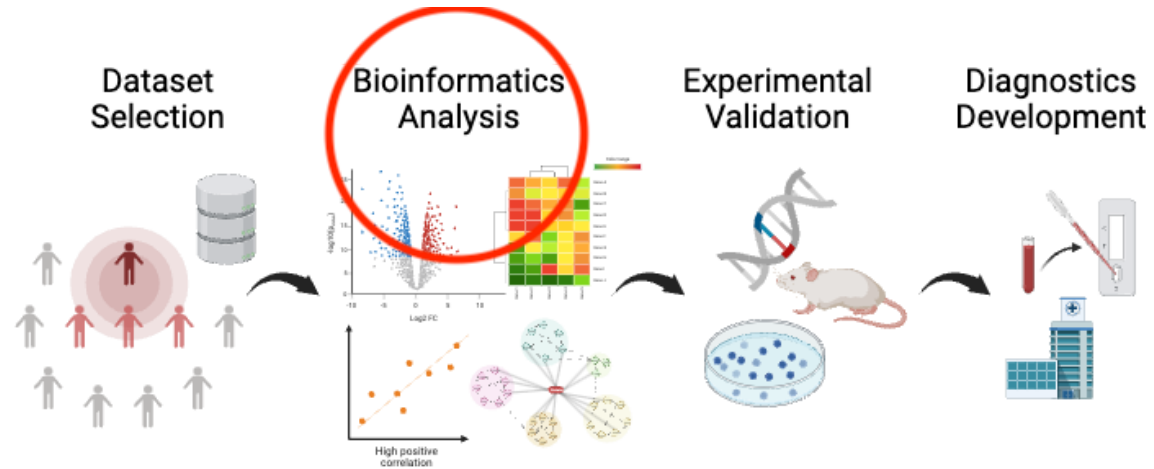
- store and manage genetic information, including DNA sequences, gene annotations, and variations in DNA
  - Database of Genomic Structural Variation (dbVar) and Database of Genotypes and Phenotypes (dbGaP)
  - Gene, GenBank, and RefSeq
  - Gene Expression Omnibus (GEO) & GEO Datasets and Genome Data Viewer (GDV)
  - International Genome Sample Resource (IGSR)
  - Cancer Single-cell Expression Map (CancerSCEM)

### How to access?

Databases provide instructions but it is usually via the following:

★ Direct download from the website    ★ FTP server    ★ API (Shell/Python/R)

## After selecting data, analysis can be done...



- via R, Python, and other programming languages

- via **Publicly-available web tools**

## **Part 2:Bioinformatics Web Tools**

# Publicly Available Web Tools

- **online** applications **accessible to a wide audience, often free/low-cost**
- aid in processing, managing, and interpreting large clinical/genomic datasets

## Main purpose

Public Web tools **turn data into interpretable results democratizing advanced data analysis without complex installations, programming skills, or high expenses**

## Advantages

- **Accessibility:** easy access to advanced data analysis capabilities for healthcare professionals and researchers
- **Cost-Effectiveness:** often free/affordable, reducing financial barriers
- **Community Support:** users benefit from a **collaborative community**, sharing knowledge and solutions, enhancing the tools' utility and troubleshooting capabilities

# Examples of Publicly available web tools

- GDC Data Portal
- cBioPortal
- UCSC Xena
- TANRIC
- miEAA
- RCoV19
- CancerSCEM analyze modules
- Integrative Genomics Viewer (IGV)
- RNAseq analysis on the Web

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart

## Harmonized Cancer Datasets

### Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

Q e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

#### Data Portal Summary [Data Release 38.0 - August 31, 2023](#)

PROJECTS 82	PRIMARY SITES 68	CASES 88,991
FILES 1,003,747	GENES 22,588	MUTATIONS 2,903,037

#### Cases by Major Primary Site

Adrenal Gland	1
Bile Duct	1
Bladder	1
Bone	1
Bone Marrow	11
Brain	1
Breast	9
Cervix	1
Colorectal	8
Esophagus	1
Eye	1
Head and Neck	1
Kidney	3
Liver	1
Lung	12
Lymph Nodes	1
Nervous System	4
Ovary	1
Pancreas	1
Pleura	1
Prostate	1
Skin	1
Soft Tissue	1
Stomach	1
Testis	1
Thymus	1
Thyroid	1
Uterus	1

#### GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

Screenshot of <https://portal.gdc.cancer.gov/>

Data Portal Website API Data Transfer Tool Documentation Data Submission Portal Publications

# How to Access and Use Web Tools

- Most explain how to access tools on their websites (web links, registration, etc.)
- A step-by-step guide (documentation) on how to use one or more of these tools for data analysis is usually provided

NIH NATIONAL CANCER INSTITUTE GDC Documentation Home API Data Portal Data Submission Data Transfer Tool Data Dictionary Data Encyclopedia

Getting Started  
Projects  
Exploration  
**Analysis**  
Generating a Cohort for Analysis  
Upload Case Set  
Upload Gene Set  
Upload Mutation Set  
Analysis Page  
Analysis Page: Set Operations  
Analysis Tab: Cohort Comparison  
Analysis Tab: Clinical Data Analysis  
Selecting a Case Set  
Enabling Clinical Variable Cards  
Exploring Clinical Card Visualizations  
Creating Custom Bins  
Other Useful Functions  
Analysis Page: Results  
Repository  
Advanced Search  
Cart and File Download  
Release Notes  
Download PDF

Analysis

Screenshot of User guide for Analysis from GDC Documentation (https://docs.gdc.cancer.gov/Data\_Portal/Users\_Guide/Custom\_Set\_Analysis)

In addition to the [Exploration Page](#), the GDC Data Portal also has features used to save and compare sets of cases, genes, and mutations. These sets can either be generated with existing filters (e.g. males with lung cancer) or through custom selection (e.g. a user-generated list of case IDs).

Note that saving a set only saves the type of entity included in the set. For example, a saved case set will not include filters that were applied to genes or mutations. Please be aware that your custom sets are deleted during each new GDC data release. You can export them and re-upload them in the "Manage Sets" link at the top right of the Portal.

### Generating a Cohort for Analysis

Cohort sets are completely customizable and can be generated for cases, genes, or mutations using the following methods:

**Apply Filters in Exploration:** Sets can be assembled using the existing filters in the Exploration page. They can be saved by choosing the "Save/Edit Case Set" button under the pie charts for case sets. This will prompt a decision to save as new case set. The same can be done for both gene and mutation filters, and can be applied and saved in the Genes and Mutations tab, respectively.

Cases (33,096) Genes (22,872) Mutations (3,142,246) OncoGrid

Primary Site Project Disease Type Gender Vital Status

Showing 1 - 20 of 33,096 cases

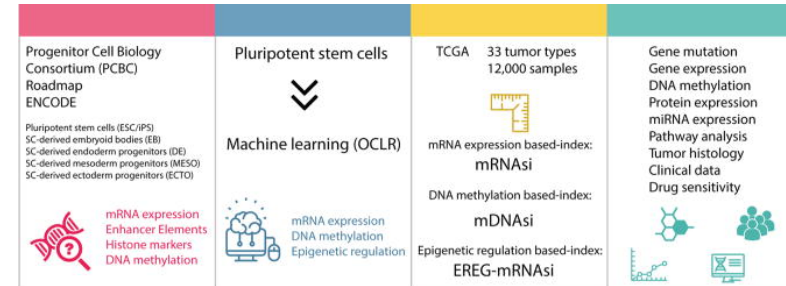
Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category							# Mut
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio	
<input type="checkbox"/> TCGA-A5-A0G2	TCGA-UCEC	Corpus uteri	Female	57	4	5	16	4	1	10	17	
<input type="checkbox"/> TCGA-EO-A22U	TCGA-UCEC	Corpus uteri	Female	56	4	5	16	4	1	10	16	26,998 12,623 (2)
<input type="checkbox"/> TCGA-FI-A2Ds	TCGA-UCEC	Corpus uteri	Female	57	4	5	16	4	1	11	16	26,139 12,482 (2)

# Case Studies – Publishing using Web tools

## ★ Case Study 1: GDC Data Portal

TCGA enabled the researchers to analyze cancer stemness in ~12,000 samples of 33 tumor types

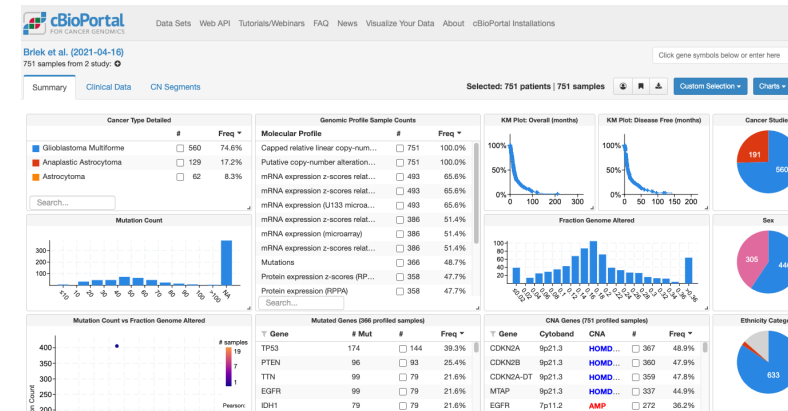
**Publication:** Fujimoto K., Ito K., Saito Y., et al. **Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation.** *Cell Reports*, 23(11), 3306-3320.e10, 2018. <https://doi.org/10.1016/j.cell.2018.03.034>



## ★ Case Study 2: cBioPortal

The researchers investigated AKT1, AKT2, AKT3, CHUK, GSK3 $\beta$ , EGFR, PTEN, and PIK3AP1 as participants of EGFR-PI3K-AKT-mTOR signaling using data from cBioPortal

**Publication:** Brlek, P.; Kafka, A.; Bukovac, A.; Pećina-Šlaus, N. **Integrative cBioPortal Analysis Revealed Molecular Mechanisms That Regulate EGFR-PI3K-AKT-mTOR Pathway in Diffuse Gliomas of the Brain.** *Cancers* 2021, 13, 3247. <https://doi.org/10.3390/cancers13133247>



# Tips & Best Practices and Pitfalls to Avoid

## DOs

- Know Your Data** – Understand the format and quality of your data
- Take your time with Data prep** – Clean and preprocess data as needed
- Select Appropriate tools** – Choose the right tool for your analysis
- Read Documentation** – Study tool guides and understand their limitations
- Pay attention to Parameters** – Set tool parameters carefully
- Record Parameters** – Keep records for reproducibility
- Validate results** – Verify results with independent data or experiments
- Secure Data** – Comply with data privacy regulations
- Seek Help** – Collaborate or ask for assistance if needed

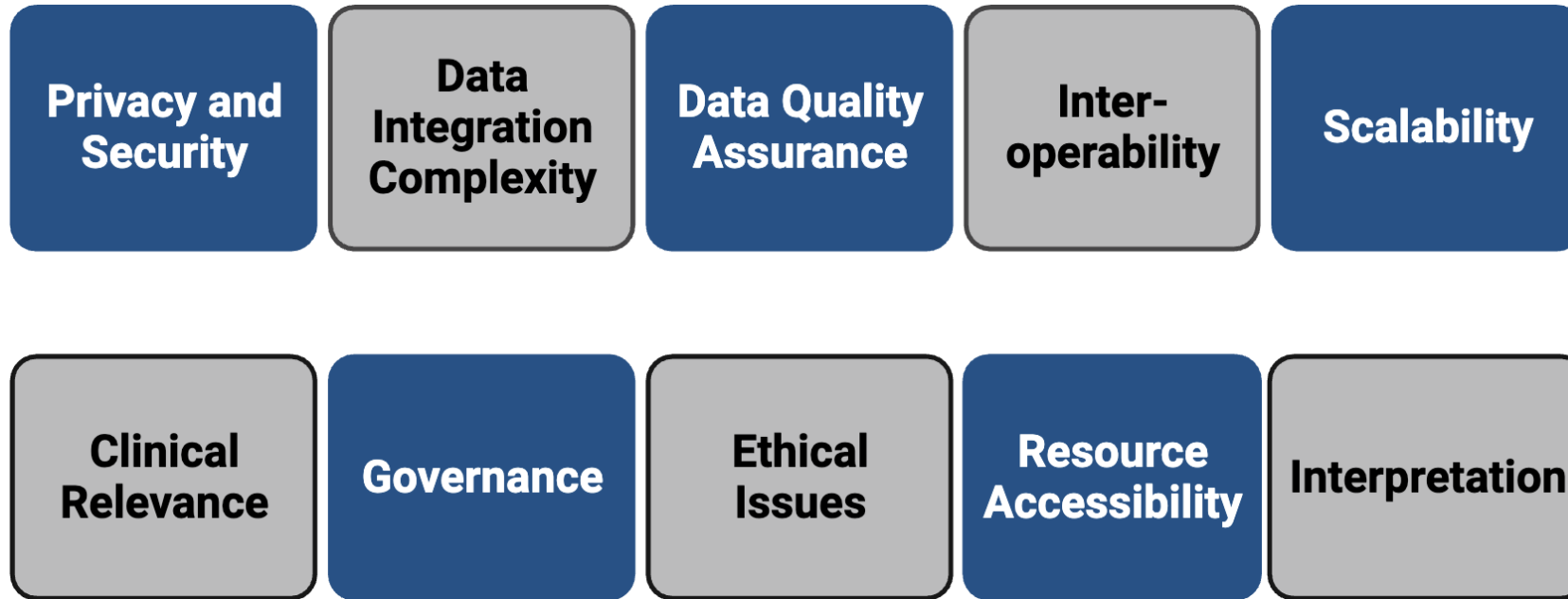
## DON'Ts

- Misinterpret your Data** – Be cautious in result interpretation
- Take Data Quality for granted** – Assess and preprocess data to ensure quality
- Use all Data when unnecessary** – Analyze relevant subsets for efficiency
- Depend on one Tool** – Use multiple tools for comprehensive analysis
- Ignore Updates** – Use the latest tool versions
- Forgo Resource Check** – Check hardware for computational capacity
- Forget Publication Quality** – Follow best practices for reporting
- Neglect Ethical Considerations** – Respect ethical guidelines and permissions



# **Part 3: Challenges and Future of Clinical Big Data Analysis**

## Key Challenges in Clinical Big Data Analysis with existing web tools



# Future of Clinical Big Data Analysis



# Collaboration and Integration

## Interdisciplinary Collaboration

Teams of healthcare providers, data scientists, and researchers working together to drive innovation



## Integration into Routine Healthcare

Seamless incorporation of data analysis into everyday healthcare practices for data-driven decision-making and personalized care



## Global Data Sharing

Enhanced collaboration and sharing of data among healthcare institutions and researchers to deepen disease understanding and improve treatments



# Current research | Project: Clinical data analyzer

迅速な臨床ビッグデータクリーニングと解析のための統合的ノーコードウェブアプリ

## Problem

- The **quality of the input data is critical** to the final results and their interpretation
- **HOWEVER**, in Healthcare and Medicine, there are many examples of rich but **unorganized, incomplete, and inconsistent data**

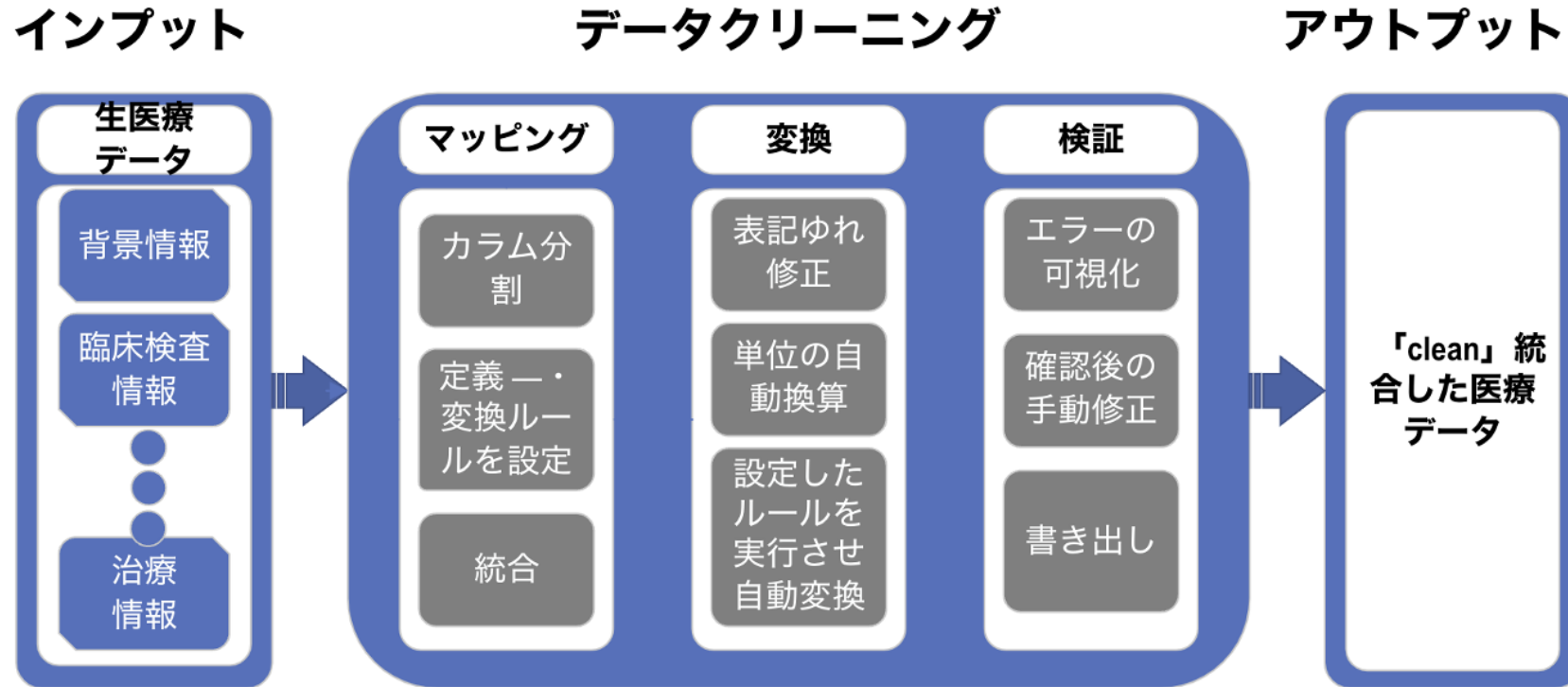
## Solution: Clinical Data Analyzer

Integrated no-code web app development for rapid clinical big data cleaning and analysis (**Collaborators wanted!**)

- a free web application that allows medical practitioners to quickly and easily construct initial hypotheses from data with so-called no-code
- consists of **three** major tools:
  1. Data cleaner
  2. Patient Finder
  3. Exploratory Analyzer



# Current research | 1. Data cleaner: a semi-automated preprocessing data cleaning tool (1)



# Current research | 1. Data cleaner: a semi-automated preprocessing data cleaning tool (2)

The screenshot shows a web browser window titled "A Web Page" with the URL <https://www.clinicaldataanalyzer-ycu.com/datacleaner>. The main content area is titled "Clinical Data cleaning" and is divided into several functional sections:

- File uploads (raw):** "Clinical data (single file)" with a "Choose" button and the filename "KidneyNHealthyMetaRaw.csv".
- Data mapping:** A table for mapping original labels to new labels:

Original label	New label
病名	Disease
年齡	Age
性別	Gender
診斷日	Diagnosis_date
...	...
組織	Subtype
- Data conversion:** Two tables showing conversion rules:

Disease	original	new
Kidney	Kidney	Malignant neoplasm of unspecified kidney, ...
Kidney cancer	Kidney cancer	Malignant neoplasm of unspecified kidney, ...
Healthy	Healthy	Encounter for general examination without complaint, ...
Non-cancer	Non-cancer	Encounter for general examination without complaint, ...

Gender	original	new
男	男	Male
女	女	Female
M	M	Male
f	f	Female

Subtype	original	new
SCC	SCC	Squamous, transitional cell carcinoma
RCC	RCC	Renal cell carcinoma
renal cell carcinoma	renal cell carcinoma	Renal cell carcinoma
carcinoma	carcinoma	Unspecified carcinoma
- Data validation:** "Conversion rule" section with buttons for Dictionary, Number (range 1-150), Date (format YYYY/MM/DD), and Dictionary.
- Factors of interest:** A list of factors: 病名, 年齡, 性別, 數值, 合併症, 吸煙, stage, 組織.
- Navigation:** A sidebar on the left contains "Data cleaner", "Patient Finder", "Exploratory Analysis".
- Output:** "Cleaned Data" section with a "Download output" button.

# Current research | 2. Patient Finder: a patient cohort selector and visualizer tool

A Web Page

https://www.clinicaldataanalyzer-ycu.com/patientfinder

### Patient finder

**File upload**  
 Group 1: Healthy.csv  
 Group 2: KidneyCancer.csv

**Disease(s)**  
 Group 1: Non-cancer, Benign  
 Group 2: Kidney

**Total (n)**  
 Group 1: 30  
 Group 2: 30

**Clinical factor (s)**  
 5 out of 30 factors selected

**Stage L1**

**Age**

I 25% 10-20 25%  
 II 25% 20-40 25%  
 III 25% 50-60 25%  
 IV 25% >60 25%

**Gender**

... Male 50%  
 Female 50%

**Patient Demography**

factor	groups	Non-cancer	Kidney
n		27	29
age	Mean (SD)	72.89 (10.15)	62.03 (11.64)
gender	female	6 (22.22%)	8 (27.59%)
	male	21 (77.78%)	21 (72.41%)
smoking history	Current use	6 (22.22%)	2 (6.897%)
	Never used	11 (40.74%)	16 (55.17%)
	Past used	10 (37.04%)	11 (37.93%)
alcohol	no	14 (51.85%)	13 (44.83%)
	yes	12 (44.44%)	16 (55.17%)

[Download patient list](#) [Download table](#)

**Smoking history**

Group	Current use	Never Used	Past used
Group 1	20	45	35
Group 2	20	20	60

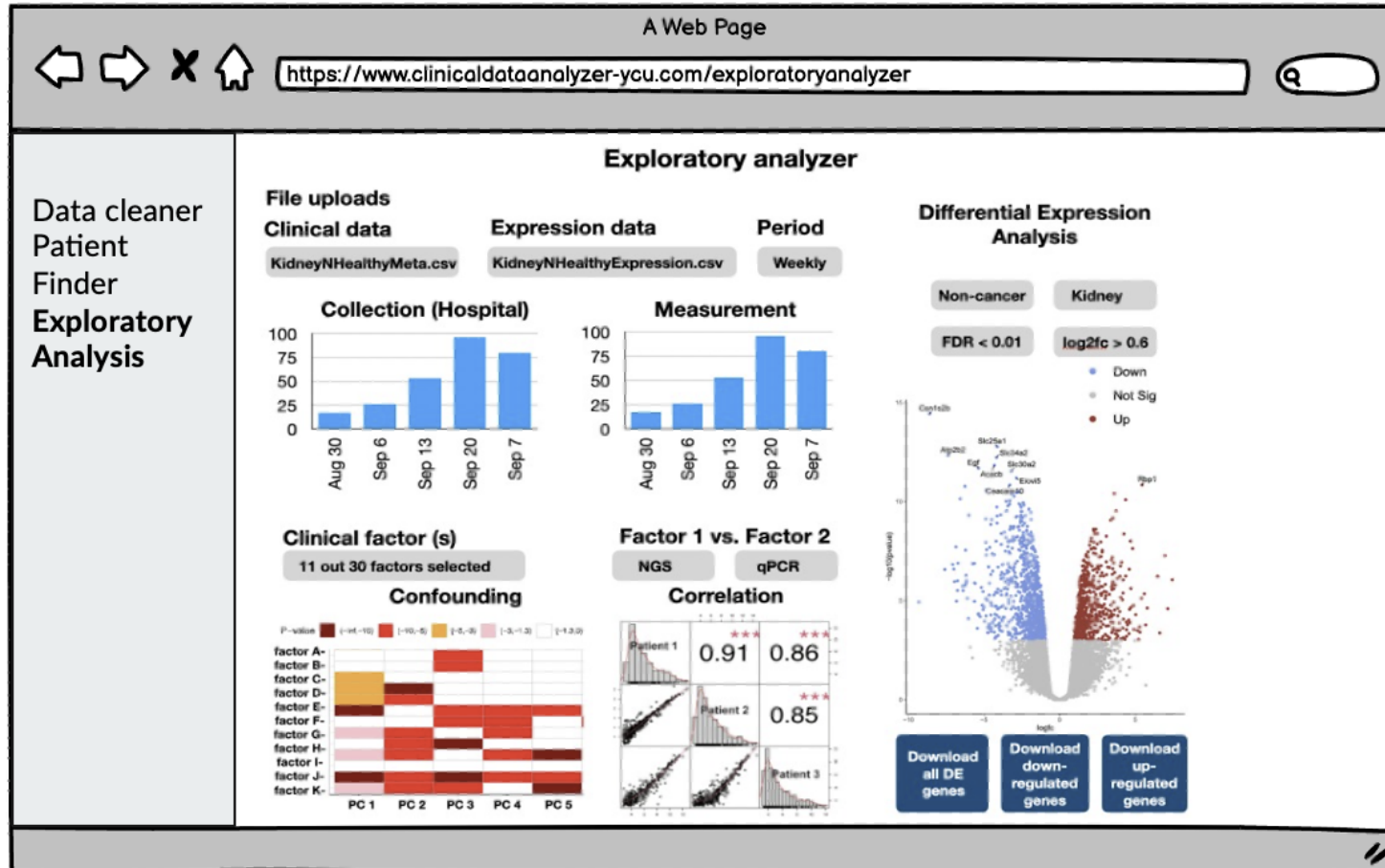
**Alcohol**

Group	Yes	No
Group 1	60	40
Group 2	55	45

Data cleaner  
 Patient Finder  
 Exploratory Analysis



# Current research | 3. Exploratory Analyzer: an interactive data analysis tool

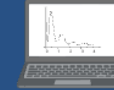


## Future research | Projects/Research Interest

Creation of databases and user-friendly web resources for biological and medical data analysis and exploration



Computational assessment of gene signatures across different cancer types



Identification of biomarkers for the diagnosis and treatment of diseases, including cancer



Employing a multi-dimensional genomic approach to enhance our comprehension of stage 0 cancers



Interested in becoming a collaborator?

Contact me: [sakura.maezono\[at\]yokohama-cu.ac.jp](mailto:sakura.maezono@yokohama-cu.ac.jp)

Contact us: [bioinfo\[at\]yokohama-cu.ac.jp](mailto:bioinfo@yokohama-cu.ac.jp)

# Summary

1. Clinical Big Data Analysis has facilitated the **extraction of valuable insights** from the continuously expanding healthcare and medical data

2. バイオインフォマティクス Web ツールは、アクセシビリティ、費用対効果、コミュニティ サポートの存在を通じて**高度なデータ分析を民主化します**。これらのツールを効果的に使用するには、正しいデータ入力、ドキュメントの読み取り、データ検証の知識が必要です。

3. **協力し、Web ツールを使用し、世界中でデータを共有することで**、病気についての理解を深め、現場のデータ品質、拡張性、ガバナンスなどの課題に対処できます。

## 持ち帰りメッセージ

バイオインフォマティクス Web ツールを活用して、増え続ける臨床ビッグデータに飛び込むことができます。

**コラボレーションとデータの共有**を通じて、**あなたは**病気に関する私たちの集合的な知識に積極的に貢献し、医療のより明るい未来を育むことができます。

# 質問とディスカッション

