

# ACROBAT - ett multi-infärgat histologiskt dataset från rutindiagnostik av bröstcancer skannat med WSI för digital patologi

**SND-ID:** 2022-190-1. **Version:** 1. **DOI:** <https://doi.org/10.48723/w728-p041>

## Ladda ner data

test.zip (68.11 GB)

train\_part1.zip (71.47 GB)

train\_part2.zip (70.59 GB)

train\_part3.zip (75.91 GB)

train\_part4.zip (71.63 GB)

train\_part5.zip (69.09 GB)

valid.zip (21.79 GB)

## Tillhörande dokumentation

df\_acrobat\_meta\_readme.txt (2.91 KB)

df\_acrobat\_meta.csv (1.11 MB)

test\_zip\_listing.txt (30.52 KB)

train\_part1\_zip\_listing.txt (35.68 KB)

train\_part2\_zip\_listing.txt (36.54 KB)

train\_part3\_zip\_listing.txt (36.17 KB)

train\_part4\_zip\_listing.txt (35.48 KB)

train\_part5\_zip\_listing.txt (36.01 KB)

valid\_zip\_listing.txt (10.06 KB)

zipfiles\_sha1\_checksums.txt (418 byte)

## Ladda ner alla filer

2022-190-1-1.zip (~448.6 GB)

## Citering

Rantalainen, M., & Hartman, J. (2023) ACROBAT - ett multi-infärgat histologiskt dataset från rutindiagnostik av bröstcancer skannat med WSI för digital patologi (Version 1) [Dataset]. Karolinska Institutet. Tillgänglig via: <https://doi.org/10.48723/w728-p041>

## Alternativ titel

ACROBAT

## Skapare/primärforskare

[Mattias Rantalainen](#) - Karolinska Institutet, Institutionen för medicinsk epidemiologi och biostatistik

[Johan Hartman](#) - Karolinska Institutet, Institutionen för onkologi-patologi

## Forskningshuvudman

## Beskrivning

ACROBAT-databasen består av 4212 mikroskopibilder (whole-slide-image, WSI) från 1153 kvinnliga primära bröstcancerpatienter. WSIs i datasetet finns tillgängliga i 10X förstoring och visar vävnadssnitt från bröstcancerresektionsprover som infärgats med hematoxylin och eosin (H&E) eller immunhistokemi (IHC). För varje patient finns en WSI av H&E-färgad vävnad och minst en och upp till fyra WSI av motsvarande vävnad som infärgats med de diagnostiska rutininfärgningarna ER, PGR, HER2 och Ki67. Datasetet skapades som en del av CHIME-studien ([chimestudy.se](http://chimestudy.se)) och dess primära syfte var att underlätta ACROBAT WSI registration challenge ([acrobot.grand-challenge.org](http://acrobot.grand-challenge.org)). De histopatologiska preparaten kommer från rutinarbetsflödet inom den diagnostiska patologin och digitaliserades för forskningsändamål vid Karolinska Institutet (Stockholm, Sverige). Skapandet av bilderna liknar det rutinmässiga arbetsflödet för digitalisering av patologibilder, med hjälp av tre olika Hamamatsu WSI-skannrar, närmare bestämt en NanoZoomer S360 och två NanoZoomer XR. WSI:erna i detta dataset åtföljs av en datatabell med en rad för varje WSI, som anger ett anonymiserat patient-ID, infärgnings- eller IHC-antikroppstypen för varje WSI, samt förstoring och mikrometer per pixel på varje tillgänglig upplösningsnivå. Automatiserad utvärdering av registreringsalgoritmers prestanda är möjlig via webbplatsen ACROBAT Challenge, baserad på över 37000 annoterade par från 13 annoterare som riktmärken. Även om det primära syftet med detta dataset var att utveckla och utvärdera WSI-registreringsmetoder, har det potential att möjliggöra forskning inom ramen för digital patologi, till exempel inom områdena infärgningsstyrd inlärning, virtuell infärgning, icke-vägled inlärning och modeller som är oberoende av färgningsmetod.

Datasetet består av tre delmängder, tränings-, validerings- och testset, baserad på ACROBAT WSI registration challenge. Det finns 750 fall i utbildningssetet, för vart och ett av fallen finns en H&E WSI och en till fyra IHC WSI:er tillgängliga, med totalt 3406 WSI:er. Valideringssetet består av 100 fall med totalt 200 WSI och testsetet av 303 fall med totalt 606 WSI. Både för validerings- och testsetet finns en H&E WSI samt en slumpmässigt utvald IHC WSI tillgänglig.

WSI:erna anonymiserades genom att de associerade makrobilderna raderats, genom att filnamn med slumpmässiga fall-ID genererats och genom att metadatafält med eventuell persondata skrivits över. Hamamatsu NDPI-filerna konverterades sedan med libvips ([libvips.org/](http://libvips.org/)). WSI:erna finns tillgängliga som generiska TIFF WSI:er ([openslide.org/formats/generic-tiff/](http://openslide.org/formats/generic-tiff/)) med 10X förstoring och lägre bildnivå.

Datasetet är tillgängligt för nedladdning i sju separata ZIP-arkiv, fem för träningsdata (train\_part1.zip (71,47 GB), train\_part2.zip (70,59 GB), train\_part3.zip (75,91 GB), train\_part4.zip (71,63 GB) och train\_part5.zip (69,09 GB)), ett för valideringsdata (valid.zip 21,79 GB) och ett för testdata (test.zip 68,11 GB).

Fillistningar och kontrollsummor i SHA1-format finns tillgängliga för att kunna kontrollera arkiv/dataintegritet vid nedladdning.

Även om det är hjälpsamt att användare meddelar SND om eventuella publikationer som använder denna datamängd genom att skicka ett e-postmeddelande till [request@snd.gu.se](mailto:request@snd.gu.se), notera att detta inte är ett krav för att använda uppgifterna.

## Data innefattar personuppgifter

Nej

## Språk

[Engelska](#)

## Analysenhet

[Individ/patient/person](#)

## Population

Anonymiserade kvinnliga patienter med primär bröstcancer, från Stockholmsregionen

## Studiedesign

Observationsstudie

## Urvalsmetod

Se beskrivningen på engelska.

## Tidsperiod(er) som undersökts

2012 - 2018

## Antal individer/objekt

1153

## Dataformat / datastruktur

[Stillbild](#)

## Datainsamling 1

- Beskrivning av insamlingsmetod: Arkiverade slides med vävnadsmaterial för klinisk rutindiagnostik skannades med hjälp av WSI-skannrar vid Karolinska Institutet.
- Tidsperiod(er) för datainsamling: 2012 - 2018
- Datainsamlare: Karolinska Institutet
- Instrument: NanoZoomer XR (Tekniskt/-a instrument) - Hamamatsu WSI-skanner
- Instrument: NanoZoomer S360 (Tekniskt/-a instrument) - Hamamatsu WSI-skanner

## Geografisk utbredning

Geografisk plats: [Stockholms län](#)

## Ansvarig institution/enhet

Institutionen för medicinsk epidemiologi och biostatistik

## Medverkande

Masi Valkonen - Åbo universitet, Institute of Biomedicine

Kimmo Kartasalo - Karolinska Institutet, Institutionen för medicinsk epidemiologi och biostatistik

Kajsa Ledesma Eriksson - Karolinska Institutet, Institutionen för medicinsk epidemiologi och biostatistik

Leena Latonen - Östra Finlands universitet, Institute of Biomedicine

Constance Boissin - Karolinska Institutet, Institutionen för medicinsk epidemiologi och biostatistik

Yanbo Feng - Karolinska Institutet, Institutionen för medicinsk epidemiologi och biostatistik  
Philippe Weitz - Karolinska Institutet, Institutionen för medicinsk epidemiologi och biostatistik  
Dusan Rasic - Sjællands universitetshospital, Patologiafdelingen  
Sonja Koivukoski - Östra Finlands universitet, Institute of Biomedicine  
Pekka Ruusuvoori - Åbo universitet, Institute of Biomedicine  
Circe Carr - Åbo universitet, Institute of Biomedicine  
Sandra Pouplier - Sjællands universitetshospital, Department of Surgical Pathology  
Leslie Solorzano - Karolinska Institutet, Institutionen för medicinsk epidemiologi och biostatistik  
Abhinav Sharma - Karolinska Institutet, Institutionen för medicinsk epidemiologi och biostatistik  
Anne-Vibeke Laenkholm - Sjællands universitetshospital, Institute of Biomedicine  
Aino Kuusela - Åbo universitet, Institute of Biomedicine

### **Finansiering 1**

- Finansiär: Vetenskapsrådet

### **Finansiering 2**

- Finansiär: ERA PerMed
- Diarienummer hos finansiär: ERAPERMED2019-224-ABCAP
- Projektnamn på ansökan: Advancing Breast Cancer histopathology towards AI-based Personalised medicine

### **Finansiering 3**

- Finansiär: Cancerfonden

### **Etikprövning**

Stockholm - dnr 2017/2106-31

Tillägg: 2018/1462-32

### **Forskningsområde**

[Vetenskap och teknologi](#) (CESSDA Topic Classification)

[Informationsteknik](#) (CESSDA Topic Classification)

[Medicinsk bildbehandling](#) (Standard för svensk indelning av forskningsämnen 2011)

[Medicin och hälsovetenskap](#) (Standard för svensk indelning av forskningsämnen 2011)

[Cancer och onkologi](#) (Standard för svensk indelning av forskningsämnen 2011)

### **Nyckelord**

[Brösttumörer](#), [Färger](#), [Eosin y](#), [Hematoxylin](#), [Datorstödd bildbehandling](#), [Immunhistokemi](#), [Klinisk patologi](#), [Färgning och märkning](#), [Wsi](#), [Image registration](#), [Whole-slide-image](#), [Digital patologi](#), [Beräkningspatologi](#), [Bröstcancer](#)

### **Publikationer**

Weitz, P. et al., (2022). ACROBAT -- a multi-stain breast cancer histological whole-slide-image data set from routine diagnostics for computational pathology. doi:10.48550/ARXIV.2211.13621

**DOI:** <https://doi.org/10.48550/ARXIV.2211.13621>

Weitz P, Valkonen M, Solorzano L, Carr C, Kartasalo K, Boissin C, Koivukoski S, Kuusela A, Rasic D, Feng Y, Sinius Pouplier S, Sharma A, Ledesma Eriksson K, Latonen L, Laenholm AV, Hartman J, Ruusuvuori P, Rantalainen M. A Multi-Stain Breast Cancer Histological Whole-Slide-Image Data Set from Routine Diagnostics. Sci Data. 2023 Aug 24;10(1):562.

**DOI:** <https://doi.org/10.1038/s41597-023-02422-6>

## Tillgänglighetsnivå

Åtkomst till data via SND  
Data är fritt tillgängliga

## Användning av data

[Att tänka på vid användning av data som delas via SND](#)

## Licens

[CC BY 4.0](#)

## Versioner

Version 1. 2023-01-02

## Hemsida

<https://acrobat.grand-challenge.org/>

<https://chimestudy.se/>

## Kontakter för frågor om data

Philippe Weitz

[philippe.weitz@ki.se](mailto:philippe.weitz@ki.se)

Mattias Rantalainen

[mattias.rantalainen@ki.se](mailto:mattias.rantalainen@ki.se)

## Denna resurs har följande relationer

Refereras till av <https://github.com/rantalainenGroup/ACROBAT>

## Ladda ner metadata

[DataCite](#)

[DDI 2.5](#)

[DDI 3.3](#)

[DCAT-AP-SE 2.0](#)

[JSON-LD](#)

[PDF](#)

[Citation \(CSL\)](#)

[Filöversikt \(CSV\)](#)

**Publicerad:** 2023-01-02

**Senast uppdaterad:** 2023-10-20