MP70-02

An overview of publicly available patient-centered prostate cancer datasets

Tim Hulsen[1], Chris Bangma[2]

[1]Philips Research, department of Professional Health Solutions & Services, Eindhoven, the Netherlands
[2]Erasmus MC, department of Urology, Rotterdam, the Netherlands

Introduction and Objectives

Prostate cancer (PCa) is the second most common cancer in men, and the second leading cause of death from cancer in men. Many studies on PCa have been carried out, each taking much time before the data is collected and ready to be analyzed. However, on the internet there is already a wide range of PCa datasets available, which could be used for data mining, predictive modelling or other purposes, reducing the need to setup new studies to collect data. In the current scientific climate, moving more and more to the analysis of 'big data' and large, international, multi-site projects, these datasets could be proven extremely valuable. This review presents an overview of publicly available patient-centered PCa datasets, divided into three categories (clinical, genomics and imaging) to enable researchers to select a suitable dataset for analysis, without having to go through days of work to find the right data.

Methods

Scientific literature databases and academic social network sites were searched to acquire a list of human PCa databases. We also used the information from other reviews, and asked experts for their input. All databases in the combined list were then checked for public availability. Only databases that were either directly publicly available or available after signing a research data agreement or retrieving a free login were selected for inclusion in this review. Data should be available to commercial parties as well. This paper focuses on patient-centered data, so the genomics data section does not include gene-centered databases or pathway-centered databases.

Results

We identified 40+ publicly available, patient-centered PCa datasets. Some of these consist of different smaller datasets. Some of them contain combinations of datasets from the three data domains: clinical data, imaging data and genomics data. Only one dataset contains information from all three domains. This review presents all datasets and their characteristics: number of subjects, clinical fields, imaging modalities, expression data, mutation data, biomarker measurements, etc.

Data source	Dataset Name	Clinical	Genomics	Imaging	# Patients
NPCR/SEER	2001-2014 Database	31 fields			2884909
NPCR/SEER	2005-2014 Database	25 fields			2092803
SEER	YR1973_2014.SEER9	136 fields			618894
SEER	YR2000_2014.CA_KY_LO_NJ_GA	136 fields			434068
SEER	YR1992_2014.SJ_LA_RG_AK	136 fields			158912
PLCO	Prostate				77000
PLCO	Prostate Screening				77000
PLCO	Prostate Screening Abnormalities				77000
PLCO	Prostate Diagnostic Procedures				77000
PLCO	Prostate Medical Complications				77000
PLCO	Prostate Treatments				77000
SEER	YR2005.LO_2ND_HALF	136 fields			1352
cBioPortal/Synapse	GENIE	11 fields	Mutation data		701
cBioPortal/ICGC/GDC/TCIA	Prostate Adenocarcinoma (TCGA, Provisional), aka PRAD-US	96 fields	Mutation data and copy number alteration data	CT, PT, MR images and tissue slide images	498
cBioPortal	Genomic Hallmarks of Prostate Adenocarcinoma (CPC-GENE)	87 fields	Comprehensive genomic profiling data		477
cBioPortal	MSK-IMPACT Clinical Sequencing Cohort (MSKCC): Prostate Cancer	12 fields	Targeted sequencing data		451
TCIA	PROSTATEx Challenge			MR (T2W, PD-W, DCE and DW) images	346
cBioPortal	Prostate Adenocarcinoma (TCGA)	85 fields	Integrated profiling data		333
cBioPortal	Prostate Adenocarcinoma (MSKCC)	18 fields	Primary and metastatic samples, cell lines and xenografts data		216
ICGC	PRAD-UK: Prostate Adenocarcinoma - United Kingdom	6 files	Mutation data (SSM, CNSM, StSM)		216
ICGC	EOPC-DE: Early Onset Prostate Cancer - Germany	6 files	Mutation data (SSM, CNSM, StSM)		211
cBioPortal	Metastatic Prostate Cancer, SU2C/PCF Dream Team	15 fields	Comprehensive analysis data		150
ICGC	PRAD-CA: Prostate Adenocarcinoma - Canada	6 files	Mutation data (SSM, CNSM, StSM, SGV, METH-A)		125
cBioPortal	Prostate Adenocarcinoma (Broad/Cornell)	10 fields	Comprehensive profiling data		112
cBioPortal	Prostate Adenocarcinoma CNA study (MSKCC)	32 fields	Copy-number profiling data		104
R ElemStatLearn package	prostate	9 fields			97
TCIA	Prostate-Diagnosis	4 fields		MR (T1, T2 and DCE) images, and segmentation data	92
cBioPortal	Neuroendocrine Prostate Cancer ((Trento/Cornell/Broad)	12 fields	Whole exome and RNAseq data		81
TCIA	Prostate-3T			MR (T2W) images, and segmentation data	64
cBioPortal	Prostate Adenocarcinoma (Fred Hutchinson CRC)	22 fields	Comprehensive profiling data		63
cBioPortal	Prostate Adenocarcinoma, Metastatic (Michigan)	21 fields	Comprehensive profiling data		59
cBioPortal	Prostate Adenocarcinoma (Broad/Cornell)	15 fields	Comprehensive profiling data		57
TCIA	Prostate Fused-MRI-Pathology			MR images, pathology images and fused Rad-Path Matlab files	28
TCIA	Prostate-MRI			MR and some PET/CT images, and pathology images	26
ICGC	PRAD-FR: Prostate Adenocarcinoma - France	6 files	Mutation data (SSM, CNSM, StSM, SGV)		25
TCIA	QIN PROSTATE			MR images	22
TCIA	NaF Prostate			PET/CT images, and DICOM metadata digest	9
cBioPortal	Prostate Adenocarcinoma Organoids (MSKCC)	13 fields	Exome profiling data		7
GEO	51 sub-datasets		Expression data		Many
ArrayExpress	117 sub-datasets		Expression data		Many

Conclusions

Despite all the attention that has been given to making this overview of publicly available databases as extensive as possible, it is very likely not complete, and will also be outdated soon. However, this review might help many PCa researchers to find suitable datasets to answer the research question with, without the need to start a new data collection project. In the coming era of big data analysis, overviews like this are becoming more and more useful.
A more detailed review paper will be published in the near future.