The Comprehensive Data Resource: An informatics system for complex, multi-omics research studies

C. R. Kinsinger,¹ L. Qi,² J. McLean,² P. HariHaran,² S. Mareedu,² K. Um,² D. Tabor,² S. Hunter,² L. Hannick,² M. Thiagarajan²

Abstract: The Comprehensive Data Resource (CDR) tracks and qualifies biospecimens for the National Cancer Institute’s Clinical Proteomic Tumor Analysis Consortium (CPTAC). The CPTAC project collects human tumor, matched non-malignant tissue, and germline DNA for comprehensive characterization by genomics, proteomics, and imaging technologies. Using Oracle 11gR2 for the backend database, the CDR captures biospecimen data through a series of forms, including: screening and eligibility, consent and enrollment, specimen procurement, pathology, molecular quality control, and clinical data. Quality control of the data is managed through automated rules and manual review by a data management team. The CDR is FISMA-moderate, which allows it to contain identifiable data, including PII and PHI. This limited dataset is made available to authorized users through dynamic content redaction. This presentation will describe the historical context, development, and future plans of the CDR.

The CPTAC Program

CPTAC Tissue Types

CPTAC code	Cancer	Percent Tumor Nuclei	Percent Total Cellularity	Percent Necrosis	Maximum tumor Ischemic Time (min)	Accrual Status
CCRC	Clear Cell Renal Carcinomas	≥80%	≥50%	≤20%	30	Complete
UCEC	Uterine Corpus Endometrial Carcinoma	≥80%	≥50%	≤20%	30	Complete
LUAD	Lung Adeno Carcinoma	≥50%	≥50%	≤20%	45	Complete
PDA	Pancreatic Ductal Adenocarcinoma	No minimum; presence of tumor	No minimum	≤20%	50	Accruing
GBM	Glioblastoma Multiforme	≥60%	≥50%	≤50%	30	Accruing
LSCC	Lung Squamous Cell Carcinoma	≥50%	≥50%	≤20%	45	Accruing
HNSCC	Head and Neck Squamous Cell Carcinoma	≥80%	≥50%	≤20%	30	Accruing
CM	Cutaneous Melanoma	≥80%	≥50%	≤20%	30	Accruing
SAR	Sarcomas	≥80%	≥50%	≤20%	30	Accruing
AML	Acute Myeloid Leukemia	≥20% blasts	N/A	N/A	N/A	Accruing

Biological questions:

- Are genomic aberrations detectable at protein level?
- What is their effect on protein function?
- Which events are drivers? Which are passengers?

Clinical questions:

- Can proteomic information provide a better molecular taxonomy of cancer?
- Can genotypic information guide protein marker development?
- Can proteomics explain why some targeted therapies do not work on 100% of patients with a given mutation?

Comprehensive Data Resource (CDR)

Objective

To meet the challenge of CPTAC requirements for collecting and tracking high quality biospecimen samples, a custom-built software system named the Comprehensive Data Resource (CDR) which was originally built for the GTEx program was further enhanced to perform and coordinate sample collection work flow, clinical data entry, case management, review and curation of study data. The CDR is a flexible and highly reliable informatics system to meet the demands of this complex and challenging initiative.

CDR Building Blocks

The CDR is an enterprise-level application built around the motto “Science First.” The technologies chosen to build the CDR facilitated a faster development time through rapid and agile software methods. The Grails framework was utilized along with other proven open source technologies. Existing Frederick National Laboratory for Cancer Research (FNLCR) database assets and infrastructure were leveraged, reducing costs and overhead. As the CPTAC project requirements evolved, the CDR’s data model and features were upgraded and enhanced. Iterative software releases were coordinated with scientific project milestones and program phases.

CDR Features : A Web application to streamline tissue collection

- ❖ Tracks cases from candidate enrollment to follow up
- ❖ Forms for collection, evaluation and documentation
- ❖ Reporting Tools & Analytics
- ❖ Pathology Review Tools
- ❖ Clinical Data harmonization
- ❖ Flexible search capabilities
- ❖ Data management and curation capabilities
- ❖ Cases approval tracking
- ❖ Deviations and quality tracking
- ❖ REST services
- ❖ Audit Log
- ❖ Active user tracing and LDAP/Database user

Process workflow

Reports & Analytics

Summary and Path forward

The CDR is a distributed web-based bioinformatics platform designed to support CPTAC operations. It manages and maintains multi-dimensional data models around each case. As an efficient case management tool capable of connecting to various remote informatics systems, the CDR could be adopted by the broader research community to standardize and streamline biospecimen accrual, sample tracking and biobanking operations.

CDR does these tasks well:

- Track prospective sample accrual
- Monitor quality of biospecimen data
- Monitor quality of clinical data
- Remind collection sites when to submit the next Case Report Form
- Report accrual status to project managers

Future features for CDR:

- Establish API with external facing data portal
- Provide status of omic characterization of a sample
- Forms for metastatic and recurrent sample types
- Forms for additional tumor types
- Forms for pediatric cancer types