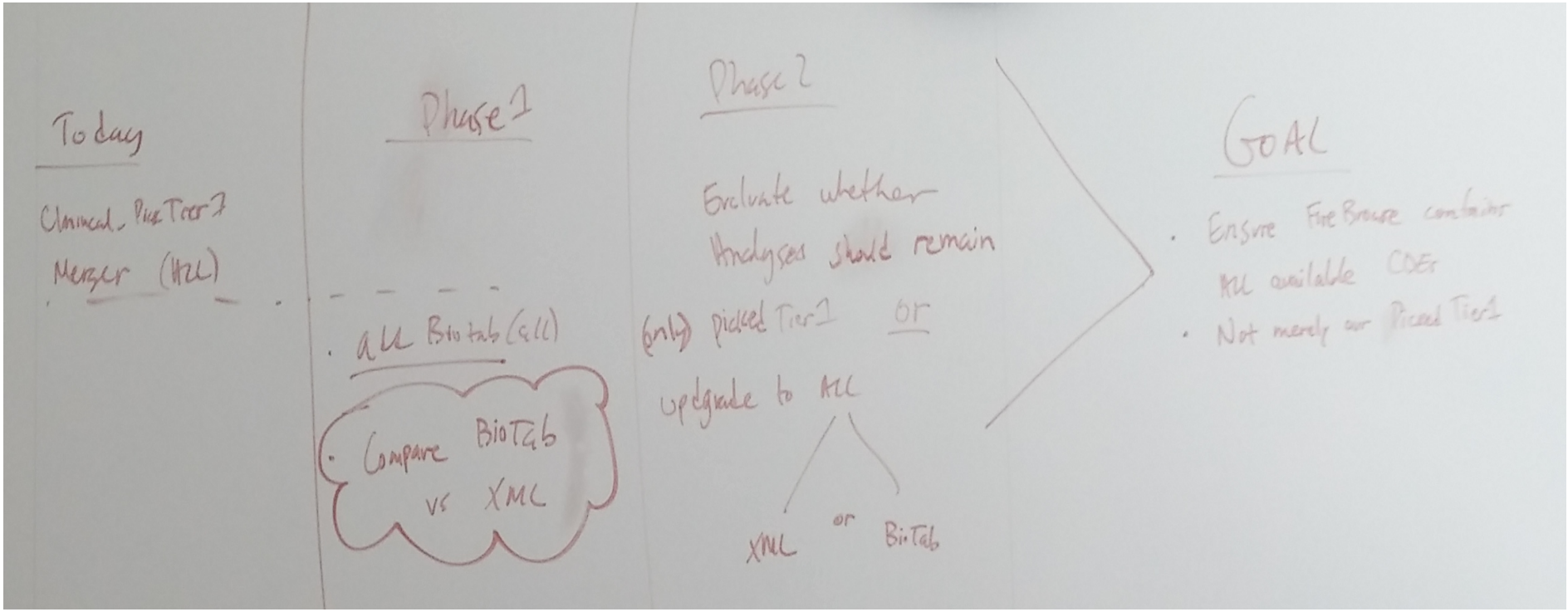
Clinical data meeting – July 21 2015



TO DO

Phase 1. Test module for BioTab(all) will go parallel with existing modules using XML

-> will compare BioTab(all) vs XML(all) in the merger output.

Phase 2. will evaluate if remain on picked or upgrade to all(of XML/BioTab)

Comparison - XML data vs BioTab data

Unique parameter count table of XML(25) and BIoTab(34) for 25 common diseases

- XML mostly has more parameters than BioTab because multiple versions are in slightly different param names in xml data.
- It is hard to obtain exactly unique parameter set to compare in xml data and the unique sample count in xml is rough number.

: BioTab clinical format has two headers.

Clinical

disease	format	xml.uniq.p.cou	biotab.uniq.p.count_h1.c	biotab.p.count_h2.c
ACC	clinical	170	164	162
BLCA	clinical	171	169	160
BRCA	clinical	204	203	194
CESC	clinical	230	244	234
CHOL	clinical	165	175	162
COAD	clinical	154	166	159
COADREAD	clinical	155	NA	NA
DLBC	clinical	155	158	184
ESCA	clinical	163	169	160
FPPP	clinical	301	NA	NA
GBM	clinical	97	115	111
GBMLGG	clinical	150	NA	NA
HNSC	clinical	164	171	161
KICH	clinical	141	136	128
KIPAN	clinical	146	NA	NA
KIRC	clinical	141	147	138
KIRP	clinical	141	150	139
LAML	clinical	166	77	78
LGG	clinical	143	161	156
LIHC	clinical	168	172	159
LUAD	clinical	151	198	188
LUSC	clinical	152	196	186
MESO	clinical	147	154	146
OV	clinical	144	125	122
PAAD	clinical	159	180	173

Auxiliary

disease	format	xml.uniq.p.coun	biotab.uniq.p.count_
CESC	auxiliary	16	12
COAD	auxiliary	15	12
COADREAD	auxiliary	15	NA
ESCA	auxiliary	15	12
HNSC	auxiliary	16	12
PAAD	auxiliary	15	12

Biospecimen

: biospecimen parameters are not proper for correlation analysis and will be excluded.

disease	format	xml.uniq.p.coun	biotab.uniq.p.count_
ACC	biospecimen	160	127
BLCA	biospecimen	168	127
BRCA	biospecimen	147	127
CESC	biospecimen	160	127
CHOL	biospecimen	166	127
COAD	biospecimen	158	127
COADREAD	biospecimen	144	NA
DLBC	biospecimen	155	127
ESCA	biospecimen	160	127
FPPP	biospecimen	137	106
GBM	biospecimen	165	127
GBMLGG	biospecimen	173	NA
HNSC	biospecimen	160	127
KICH	biospecimen	158	127
KIPAN	biospecimen	166	NA
KIRC	biospecimen	158	127
KIRP	biospecimen	166	127
LAML	biospecimen	166	83
LGG	biospecimen	171	127
LIHC	biospecimen	166	127
LUAD	biospecimen	165	127
LUSC	biospecimen	164	127
MESO	biospecimen	157	127
OV	biospecimen	160	127
PAAD	biospecimen	164	127

Ex. For the variable, "**additional_pharmaceutical_therapy**" (in SARC-TP)

In XML data

: It is hard to sort the multiple data in chronological order

```
 [1] "patient.follow_ups.follow_up-2.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy"
 [2] "patient.follow_ups.follow_up-3.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy"
 [3] "patient.follow_ups.follow_up-4.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy"
 [4] "patient.follow_ups.follow_up-5.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy"
 [5] "patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-2.additional_pharmaceutical_therapy"
 [6] "patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-3.additional_pharmaceutical_therapy"
 [7] "patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-4.additional_pharmaceutical_therapy"
 [8] "patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-5.additional_pharmaceutical_therapy"
 [9] "patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-6.additional_pharmaceutical_therapy"
[10] "patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-7.additional_pharmaceutical_therapy"
[11] "patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-8.additional_pharmaceutical_therapy"
[12] "patient.follow_ups.follow_up.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy"
[13] "patient.new_tumor_events.new_tumor_event-2.additional_pharmaceutical_therapy"
[14] "patient.new_tumor_events.new_tumor_event-3.additional_pharmaceutical_therapy"
[15] "patient.new_tumor_events.new_tumor_event-4.additional_pharmaceutical_therapy"
[16] "patient.new_tumor_events.new_tumor_event-5.additional_pharmaceutical_therapy"
[17] "patient.new_tumor_events.new_tumor_event-6.additional_pharmaceutical_therapy"
[18] "patient.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy"
```

In BioTab data

: there is no multiple versions and two CDE groups(*.nte.txt) have the parameter.

nationwidechildrens.org_clinical_follow_up_v4.0_nte_sarc.txt

```
bash:cga02:/xchip/cga_home/jcho/GDAC/clinicalData/biotab/biotab_data/sarc/bcr/biotab/clin 1207 $ cat nationwidechildrens.org_clinica
nte_sarc.txt | head -n 2
bcr_patient_uuid        bcr_patient_barcode     bcr_followup_barcode    new_tumor_event_type    new_tumor_event_site    new_tumor_ev
ent_site_other  new_tumor_event_dx_days_to      new_tumor_event_surgery new_tumor_event_surgery_days_to nte_lesion_radiologic_length
        nte_lesion_radiologic_width     nte_lesion_radiologic_depth     nte_lesion_pathologic_length    nte_lesion_pathologic_widthn
te_lesion_pathologic_depth      nte_well_or_dedifferentiated_indicator  new_tu
al_tx   discontiguous_lesion_count      pathologic_tumor_burden radiologic_tum
_status tumor_multifocal
bcr_patient_uuid        bcr_patient_barcode     bcr_followup_barcode    new_ne
ic_site new_neoplasm_occurrence_anatomic_site_text      days_to_new_tumor_ever
urgery_procedure        days_to_new_tumor_event_additional_surgery_procedure
ologic_tumor_depth      pathologic_tumor_length pathologic_tumor_width  pathol
dditional_radiation_therapy     additional_pharmaceutical_therapy       discor
ologic_tumor_burden     residual_disease_post_new_tumor_event_margin_status
```

nationwidechildrens.org_clinical_nte_sarc.txt

```
bash:cga02:/xchip/cga_home/jcho/GDAC/clinicalData/biotab/biotab_data/sarc/bcr/
head -n 2c.txt |
bcr_patient_uuid        bcr_patient_barcode     disease_multifocal_indicator
gic     tumor_burden_pathologic new_tumor_event_type    new_tumor_event_site
ays_to  new_tumor_event_surgery new_tumor_event_surgery_days_to nte_lesion_rad
lesion_radiologic_depth nte_lesion_pathologic_length    nte_lesion_pathologic_
dedifferentiated_indicator      new_tumor_event_radiation_tx    new_tumor_even
mor_event_margin_status
bcr_patient_uuid        bcr_patient_barcode     tumor_multifocal        discontiguous_lesion_count      radiologic_tumor_burden path
ologic_tumor_burden     new_neoplasm_event_type new_neoplasm_event_occurrence_anatomic_site     new_neoplasm_occurrence_anatomic_sit
e_text  days_to_new_tumor_event_after_initial_treatment new_tumor_event_additional_surgery_procedure    days_to_new_tumor_event_addi
tional_surgery_procedure        radiologic_tumor_length radiologic_tumor_width  radiologic_tumor_depth  pathologic_tumor_length path
ologic_tumor_width      pathologic_tumor_depth  new_tumor_cellular_differentiation      additional_radiation_therapy    additional_p
harmaceutical_therapy   residual_disease_post_new_tumor_event_margin_status
```

```
nationwidechildrens.org_biospecimen_aliquot_sarc.txt
nationwidechildrens.org_biospecimen_analyte_sarc.txt
nationwidechildrens.org_biospecimen_cqcf_sarc.txt
nationwidechildrens.org_biospecimen_diagnostic_slides_sarc.txt
nationwidechildrens.org_biospecimen_normal_control_sarc.txt
nationwidechildrens.org_biospecimen_portion_sarc.txt
nationwidechildrens.org_biospecimen_protocol_sarc.txt
nationwidechildrens.org_biospecimen_sample_sarc.txt
nationwidechildrens.org_biospecimen_shipment_portion_sarc.txt
nationwidechildrens.org_biospecimen_slide_sarc.txt
nationwidechildrens.org_biospecimen_tumor_sample_sarc.txt
nationwidechildrens.org_clinical_cqcf_sarc.txt
nationwidechildrens.org_clinical_drug_sarc.txt
nationwidechildrens.org_clinical_follow_up_v4.0_nte_sarc.txt
nationwidechildrens.org_clinical_follow_up_v4.0_sarc.txt
nationwidechildrens.org_clinical_nte_sarc.txt
nationwidechildrens.org_clinical_omf_v4.0_sarc.txt
nationwidechildrens.org_clinical_patient_sarc.txt
nationwidechildrens.org_clinical_radiation_sarc.txt
```

In BioTab data

: the *nte (new tumor event) CDE groups have multiple 'additional pharmaceutical_therapy'.

nationwidechildrens.org_clinical_follow_up_v4.0_nte_sarc.txt

	A	B	C	D	E	R
1	bcr_patient_	bcr_patient_barcode	bcr_followup_barcode	new_tumor_event_type	new_tumor_event_site	new_tumor_event_pharmaceutical_tx
2	bcr_patient_	bcr_patient_barcode	bcr_followup_barcode	new_neoplasm_event_type	new_neoplasm_event_occurrence_	additional_pharmaceutical_therapy
3	CDE_ID:	CDE_ID:2673794	CDE_ID:	CDE_ID:3119721	CDE_ID:3108271	CDE_ID:3427616
4	34C39A6B-D	TCGA-3B-A9HI	TCGA-3B-A9HI-F69153	Locoregional Recurrence	[Not Available]	NO
5	34C39A6B-D	TCGA-3B-A9HI	TCGA-3B-A9HI-F69153	Locoregional Recurrence	[Not Available]	YES
6	34C39A6B-D	TCGA-3B-A9HI	TCGA-3B-A9HI-F69153	Distant Metastasis	Lung	YES
7	34C39A6B-D	TCGA-3B-A9HI	TCGA-3B-A9HI-F69153	Distant Metastasis	Other, specify	YES
8	F6509C51-3E	TCGA-3B-A9HO	TCGA-3B-A9HO-F69263	New Primary Tumor	Other, specify	NO
9	BFB21F9B-31	TCGA-3B-A9HR	TCGA-3B-A9HR-F69270	Distant Metastasis	Lung	YES
10	BFB21F9B-31	TCGA-3B-A9HR	TCGA-3B-A9HR-F69270	Distant Metastasis	Liver	YES
11	BFB21F9B-31	TCGA-3B-A9HR	TCGA-3B-A9HR-F69270	Distant Metastasis	Other, specify	YES

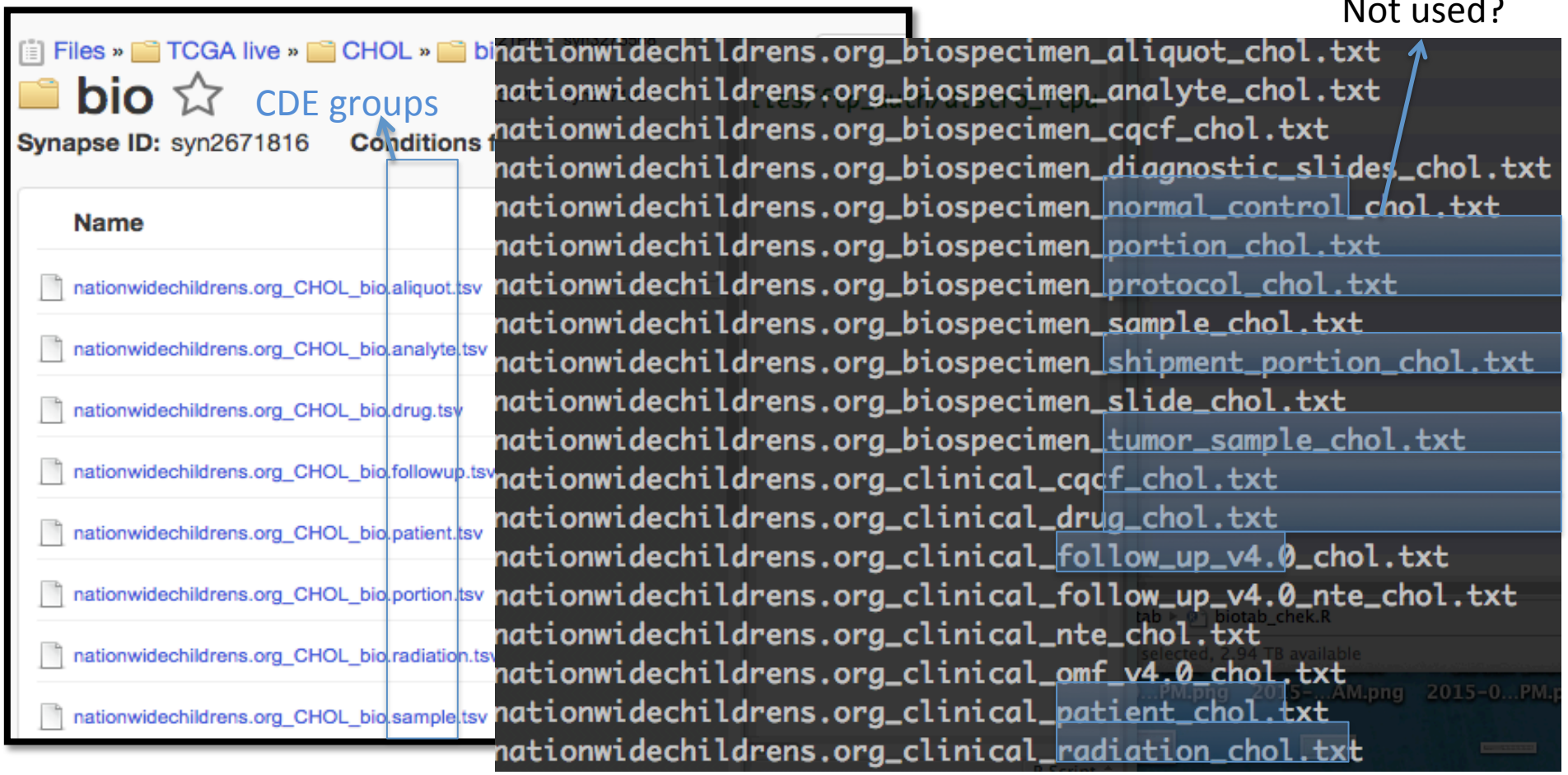
nationwidechildrens.org_clinical_nte_sarc.txt

	A	B	C	D	E	S	T	U	
1	bcr_patient_	bcr_patient_barcode	disease_mul	discontinuou	tumor_burde	nte_well_or_	new_tumor_	new_tumor_event_pharmaceutical_tx	res
2	bcr_patient_	bcr_patient_barcode	tumor_multi	discontiguou	radiologic_tu	new_tumor_	additional_ra	additional_pharmaceutical_therapy	res
3	CDE_ID:	CDE_ID:2673794	CDE_ID:6435	CDE_ID:3162	CDE_ID:3162	CDE_ID:3194	CDE_ID:3427	CDE_ID:3427616	CD
4	34C39A6B-D	TCGA-3B-A9HI	NO	[Not Availabl	[Not Availabl	De-Differenti	NO	NO	RO
5	34C39A6B-D	TCGA-3B-A9HI	[Not Availabl	[Not Availabl	[Not Availabl	[Not Availabl	NO	YES	[N
6	34C39A6B-D	TCGA-3B-A9HI	[Not Availabl	[Not Availabl	[Not Availabl	[Not Availabl	NO	YES	[N
7	34C39A6B-D	TCGA-3B-A9HI	[Not Availabl	[Not Availabl	[Not Availabl	[Not Availabl	NO	YES	[N
8	AAF93A92-2I	TCGA-3B-A9HL	YES	[Not Availabl	[Not Availabl	Well-Differer	NO	YES	[N
9	AAF93A92-2I	TCGA-3B-A9HL	YES	[Not Availabl	[Not Availabl	De-Differenti	NO	YES	[N
10	AAF93A92-2I	TCGA-3B-A9HL	NO	[Not Availabl	[Not Availabl	De-Differenti	NO	YES	[N
11	F6509C51-3E	TCGA-3B-A9HO	NO	[Not Availabl	[Not Availabl	[Not Availabl	NO	NO	RO

How to use BioTab data

Biotab data usage in CHOL AWG

: They use only few CDE groups for analysis.

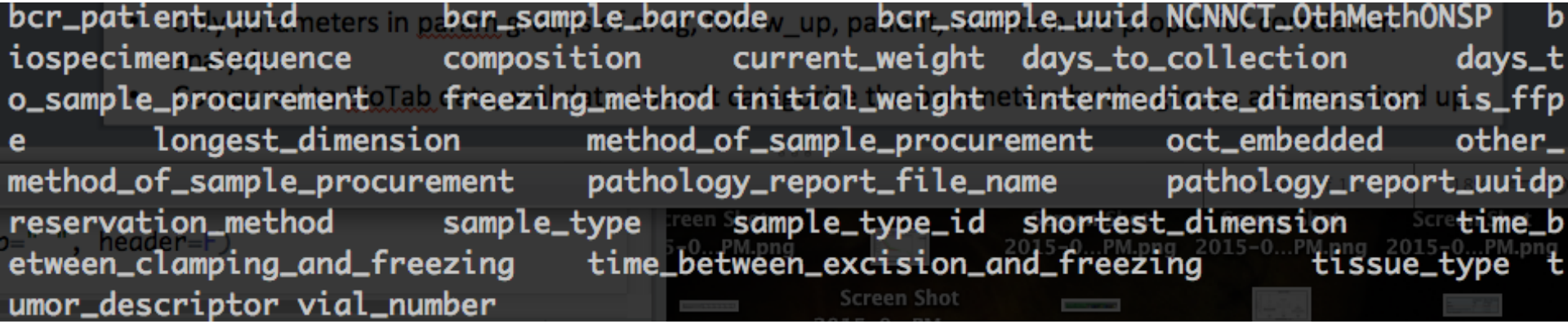


: For the picker pipeline,

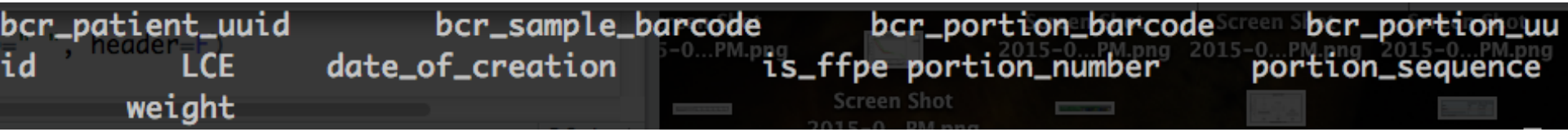
- only parameters in CDE groups of drug, follow_up, patient, radiation are proper for correlation analysis.
- Compared to BioTab data, xml data doesn't categorize the parameters by the groups and are mixed up.

Params in CDE groups of sample, portion, aliquot, analyte: not proper for correlation analysis

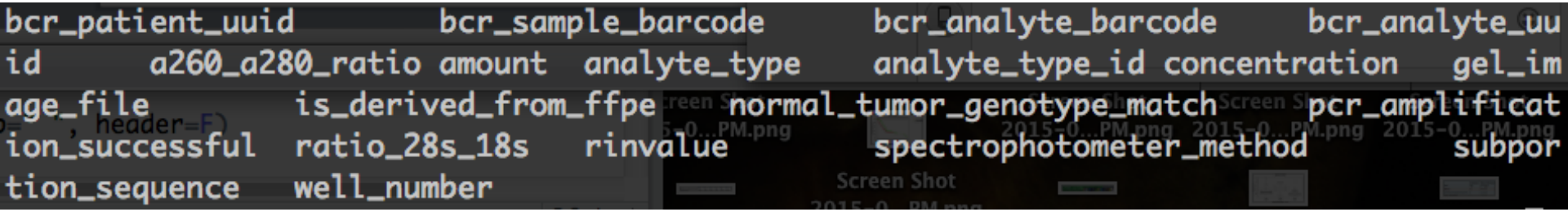
Sample



portion



aliquot



analyte

BioTab format table for all diseases

- BioTab data files saved in .txt by different parameter groups.
- CDE groups in yellow color are mostly useful for correlation analysis.
- XML data parameters are not grouped and mixed up so providing all CDEs leads meaningless parameters to be used in correlation analysis.
- Only parameters in CDE groups of drug, follow_up, patient, radiation are proper for correlation analysis.

Total 18 CDE groups

Biospecimen

Clinical

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ
1		acc	blca	blnp	blp	brca	cesc	chol	cntl	coad	dlbc	esca	fppp	gbm	hnsc	kich	kirc	kirp	laml	lcll	lcml	lgg	lihc	lndlbc	lnnh	luad	lusc	meso	misc	ov	paad	pcpg	prad	read	sarc	skcm	stad	tgct	thca	thym	ucec	ucs	uvm
2	aliquot	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	analyte	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	cqcf	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	diagnostic_slides	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	normal_control	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	portion	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	protocol	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	sample	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	shipment_portion	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	slide	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	tumor_sample	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	drug	1	1	0	0	1	1	1	0	1	0	1	0	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	nte	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	omf_v4.0	1	1	0	0	1	1	1	0	1	0	1	0	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	patient	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	radiation	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	follow_up	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	control	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

OPS meeting agenda for clinical update - 2015

June 15, 2015

1. Welcome Tim
2. iCoMut for RAS pathway: status?
    a. check integration–at present there are aggregate failures
3. iCoMut for SARC run: presentation deferred until next week, but still have legwork to do
4. Review & parcel out unanswered GDAC list questions
5. Review CompBio projects:
    a. Update clinical pipelines to potentially use BioTAB, and reveal ALL clinical parameters--not just picked.
    b. Recent / Upcoming AWG runs:
        i. PAAD:
        ii. CHOL:
    c. Bayesian NMF
    d. MGH: horizon?
    e. More?
6. Revisit prioritizing next 6-12 month dev window

f. Where would Sam fit?

5. viewGene widget RFEs:

June 1, 2015

lisease cohorts with no RSEM values

6. iCoMut for RAS pathway
7. Clinical data:
    a. why continue to use XML over biotab?
    b. Per Gordon: XML was preferred over biotab because
        i. it's the source data
        ii. from which biotab was generated (by DCC)
        iii. and biotab was more buggy
        iv. and it comes in packages of 5-6 files per cohort, which was harder to merge
    c. we need to provide ALL of clinical, not just our picked subset

Clinical data update agenda - 2015

- All clinical features
  - current xml data
  - BioTab data

  ⇒ Will try BioTab first and then will compare with xml.
- Up-to-date data
  - There are more parameters in followup data versioned (see slide 7)

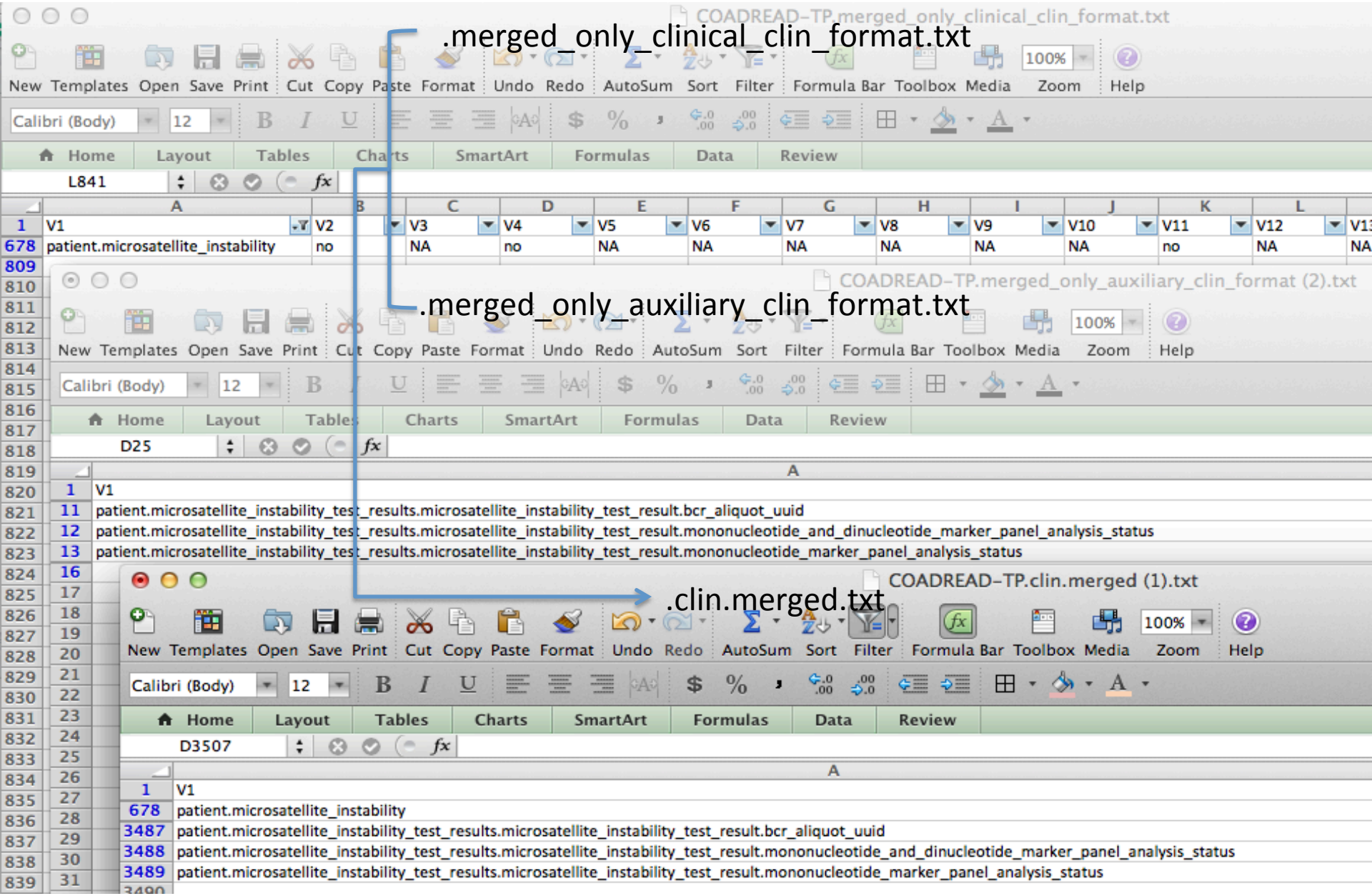
  ⇒ Will update pipeline, SelectionFileGenerator for all followup parameters such as radiation_therapy.
- BioTab data
  - XML: 3 txt files generated from the dicer running XML parser
  - BioTab: Can BioTab also be generated from the dicer? (Not yet)

  => Will check if BioTab can be added in a parameter in Clinical_Merger.
- MSI data
  - According to Gordon's email (see next slide 4), there are CDE groups unprocessed having MSI data.

  => Will skipped them as our dicer doesn't work for them.

> Gordon recently reported that he found new data type having MSI info and the data type was not processed due to ingestor failures during the 2014_10_06 GDAC ops meeting.
> According to Gordon, the new data types are as below.
> The files live here:
> 3:24pm gsaksena@voncotator2-dev /xchip/gdac_data/dcc_mirror3/platform_link $ ls -1 *micro*
> coad__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__anonymous@
> coad__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__tcga4yeo@
> read__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__anonymous@
> read__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__tcga4yeo@
> stad__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__anonymous@
> stad__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__tcga4yeo@
> ucec__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__anonymous@
> ucec__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__tcga4yeo@
> ucs__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__anonymous@
> ucs__cgcc__nationwidechildrens.org__microsat_i__fragment_analysis__tcga4yeo@
> The corresponding unprocessed flags from the dicer are here, which indicates the ingester is ignoring this data type:
> 3:27pm gsaksena@cga-cdkn2a /xchip/gdac_data/normalized/sdrf_cache $ ls -1 *unprocessed
> nationwidechildrens.org_COAD.microsat_i.mage-tab.1.6.0.txt.unprocessed
> nationwidechildrens.org_READ.microsat_i.mage-tab.1.7.0.txt.unprocessed
> nationwidechildrens.org_STAD.microsat_i.mage-tab.1.0.0.txt.unprocessed
> nationwidechildrens.org_UCEC.microsat_i.mage-tab.1.6.0.txt.unprocessed
> nationwidechildrens.org_UCS.microsat_i.mage-tab.1.0.0.txt.unprocessed

(Current) MSI data obtained from the Merge_Clinical in stddata run



Parameters

Parameter Name	Parameter Expression	Mode	Default Value
sdrfName		Simple_Expression	
doNotCreateManifest	-m	Literal	
samplestamp*	samplestamp	Simple_Expression	
annotationids*	clin__bio__nationwidechil	Literal	
datapathsfile1*	samples.choose(["clin__bi	Complex_Expression	
inFile_clinical_clin*	samples.choose(["clin__bi	Complex_Expression	
inFile_biospecimen_clin	samples.choose(["clin__bi	Complex_Expression	
inFile_auxiliary_clin	samples.choose(["clin__bi	Complex_Expression	
outPrefix*	sample_set_id	Simple_Expression	

Additional parameters?

XML data - **'Vital_status parameters'** in the PAAD-TP.clin.merged.txt

	A	B	C	D	E	F	G
1	patient.bcr_patient_barcode	tcga-2j-aab1	tcga-2j-aab4	tcga-2j-aab6	tcga-2j-aab8	tcga-2j-aab9	tcga-2j-aaba
2	patient.vital_status	dead	alive	dead	alive	dead	dead
3	patient.follow_ups.follow_up.vital_status	NA	alive	NA	alive	NA	dead
4	patient.follow_ups.follow_up-2.vital_status	NA	NA	NA	NA	NA	NA
5	patient.follow_ups.follow_up-3.vital_status	NA	NA	NA	NA	NA	NA

Final values for vital_status	**Dead**	**Alive**	**Dead**	**Alive**	**Dead**	**dead**
Where the value comes from	Primary param	Followup version1	Primary param	Followup version1	Primary param	Followup version1

- As seen in the table above,

1. Using the latest followup version 3 is not correct
2. Using only one of versions is not correct
3. NA in the latest version followup data means there is no update from the previous version.
4. Thus, data values should be the combination of all versions.
5. Also, It is not helpful to use the DCC upload date parameter to pick up the latest followup because the values are NA if there is no change from the previous version.

=> For the AWG run generator, recommend to check if there are new version of followup data appeared in the .clin.merged.txt

XML data - **'Radiaton_therapy parameters'** in the COADREAD-TP.clin.merged.txt

1	V1
565	patient.follow_ups.follow_up-2.new_tumor_events.new_tumor_event.additional_radiation_therapy
576	patient.follow_ups.follow_up-2.radiation_therapy
589	patient.follow_ups.follow_up-3.new_tumor_events.new_tumor_event.additional_radiation_therapy
600	patient.follow_ups.follow_up-3.radiation_therapy
613	patient.follow_ups.follow_up-4.new_tumor_events.new_tumor_event.additional_radiation_therapy
624	patient.follow_ups.follow_up-4.radiation_therapy
637	patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-2.additional_radiation_therapy
645	patient.follow_ups.follow_up.new_tumor_events.new_tumor_event.additional_radiation_therapy
656	patient.follow_ups.follow_up.radiation_therapy
681	patient.new_tumor_events.new_tumor_event.additional_radiation_therapy
704	patient.radiation_therapy
711	patient.radiations.radiation-2.days_to_radiation_therapy_end
712	patient.radiations.radiation-2.days_to_radiation_therapy_start
729	patient.radiations.radiation-3.days_to_radiation_therapy_end
730	patient.radiations.radiation-3.days_to_radiation_therapy_start
747	patient.radiations.radiation-4.days_to_radiation_therapy_end
748	patient.radiations.radiation-4.days_to_radiation_therapy_start
765	patient.radiations.radiation.days_to_radiation_therapy_end
766	patient.radiations.radiation.days_to_radiation_therapy_start

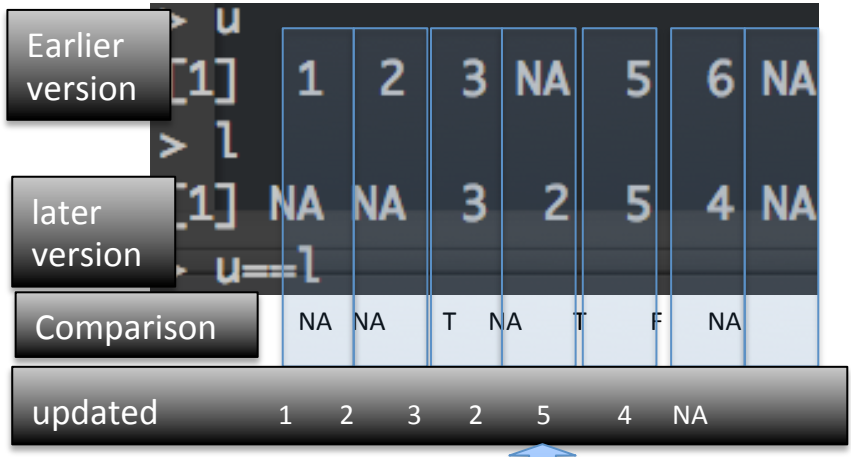
- Here, followup version 4 is the latest version but the version number is not chronological order. Our SelectionFileGenerator and All_CDEs_Parser check all values of each parameter and assign the latest value.

XML data - "**additional_phamaceutical_therapy**"

```
#** CASE1. How to order ?? **#
 patient.follow_ups.follow_up-2.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy
 patient.follow_ups.follow_up-3.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy
 patient.follow_ups.follow_up-4.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy
 patient.follow_ups.follow_up.new_tumor_events.new_tumor_event-2.additional_pharmaceutical_therapy
 patient.follow_ups.follow_up.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy
 patient.new_tumor_events.new_tumor_event.additional_pharmaceutical_therapy

Is there follow_up nodes? yes
-> find the follow_up nodes
1. find  a node not matched with "follow_up"
2. find matches with ".follow_up."
  * Is there multiple matches?
   -> find the uppderNdoe
  2.1. find matches with ".upperNode."
  2.2. find matches with ".upperNode-#."
3. find matches with ".follow_up-#."
```

After setting order, do updating across all cases



If F: override
If(earlier is NA) : override
else remain

XML data - "**days_to_radiation_therapy_start**"

```
#** CASE2. How to order ?? ** #
patient.radiations.radiation-2.days_to_radiation_therapy_start
patient.radiations.radiation-3.days_to_radiation_therapy_start
patient.radiations.radiation-4.days_to_radiation_therapy_start
patient.radiations.radiation.days_to_radiation_therapy_start

Is there follow_up nodes? no
-> find the upperNode
1. fidn matches with ".upperNode."
2. find matches with ".upperNode-#."
```