**Diagnosis date for lung cancer**

1. PreProcess path report (split path report):

perl preProcessingNew\_1.pl path path\_inter/

output:

number\_noteType\_2002-06-20.txt\_1

number\_noteType\_2002-06-20.txt\_2

1. Medtagger—Path\_inter report
2. delete empty files in output\_medtagger folder:

find . -size 0 -delete

find path\_inter -size 0 -print0 |xargs -0 rm

1. concatenate all files in output\_medtagger folder:

find . -type f -exec cat {} > ../path\_medtagger\_results  \;

1. perl lungBiopsyDiagnosisDate.pl /path\_inter  path\_medtagger\_results > diagnosis\_date/path\_diagnosis\_date
2. perl compareBiopsyDatesEGLC.pl path\_diagnosis\_date having\_dxdts.txt > no\_match
3. perl outsideReview.pl ../path\_medtagger\_results> outsideReview\_diagnosis\_date

1. MedTime
2. perl map.pl out\_outside/outsideReview\_diagnosis\_date.ann outside/outsideReview\_diagnosis\_date  > outsideReview\_diagnosis\_date\_norm
3. perl compareBiopsyDatesEGLC2.pl outsideReview\_diagnosis\_date\_norm no\_match

Questions:

1. Currently only first diagnosis date?
2. Distinctive keywords between lung cancer path report and other types of path reports
3. How many patients have outside materials in PDF?
4. Issues in CN----Add “large cell cancer”?

Cerebellum, left suboccipital #2, biopsy: Metastatic non-small \par cell carcinoma.

Thoughts:

* Add rule in future
* Add “atypical epithelial cells”, SBRT treatment, then adenocarcinoma to confirm primary lung cancer.
* Use notes to identify “recurrence” of second diagnosis, so should remove for survival analysis.
* Need to extract different dates for different primaries.
* Match for 15 of same month to my data:
* Add rule in future
* Add “atypical epithelial cells”, SBRT treatment, then adenocarcinoma to confirm primary lung cancer.
* Use notes to identify “recurrence” of second diagnosis, so should remove for survival analysis.
* Need to extract different dates for different primaries.
* Match for 15 of same month to my data:
* Inference rule:

If identified as “large cell” later, then use the date earlier associated with “large cell”.

* Add Retroperitoneum as keyword.