SPECIALIZED NAMED ENTITY RECOGNITION

FOR BREAST CANCER SUBTYPING

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Grif Hawblitzel

June 2022

COMMITTEE MEMBERSHIP

TITLE:	Specialized Named Entity Recognition for Breast Cancer Subtyping
AUTHOR:	Grif Hawblitzel
DATE SUBMITTED:	June 2022
COMMITTEE CHAIR:	Paul Anderson, Ph.D. Professor of Computer Science
COMMITTEE MEMBER:	Jean Davidson, Ph.D. Professor of Biology
COMMITTEE MEMBER:	Dongfeng Fang, Ph.D. Professor of Computer Science

ABSTRACT

Specialized Named Entity Recognition for Breast Cancer Subtyping

Grif Hawblitzel

The amount of data and analysis being published and archived in the biomedical research community is more than can feasibly be sifted through manually, which limits the information an individual or small group can synthesize and integrate into their own research. This presents an opportunity for using automated methods, including Natural Language Processing (NLP), to extract important information from text on various topics. Named Entity Recognition (NER), is one way to automate knowledge extraction of raw text. NER is defined as the task of identifying named entities from text using labels such as people, dates, locations, diseases, and proteins. There are several NLP tools that are designed for entity recognition, but rely on large established corpus for training data. Biomedical research has the potential to guide diagnostic and therapeutic decisions, yet the overwhelming density of publications acts as a barrier to getting these results into a clinical setting. An exceptional example of this is the field of breast cancer biology where over 2 million people are diagnosed worldwide every year and billions of dollars are spent on research. Breast cancer biology literature and research relies on a highly specific domain with unique language and vocabulary, and therefore requires specialized NLP tools which can generate biologically meaningful results. This thesis presents a novel annotation tool, that is optimized for quickly creating training data for spaCy pipelines as well as exploring the viability of said data for analyzing papers with automated processing. Custom pipelines trained on these annotations are shown to be able to recognize custom entities at levels comparable to large corpus based recognition.

ACKNOWLEDGMENTS

Thanks to:

- Bioinformatics Research Group, for all sorts of help.
- Andrew Guenther, for uploading this template

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1

INTRODUCTION

Breast cancer is a heterogeneous disease in which several etiologies can result in varying manifestations of disease in patients[32]. Molecular classification using gene expression and traditional pathology-driven techniques has classified breast cancer into the following five distinct molecular subtypes: luminal A, luminal B, HER-2 enriched, basal-like (triple-negative), and normal [50]. Among the subtypes, there are differences in disease progression, tissue origin, and treatment that result in prognostic significance [39][46]. The ability to clinically diagnose the subtypes as soon as possible in the identification process can help clinicians in defining prognosis and treatment strategies, to improve patient care outcomes [6]. Breast cancer is one of the most researched diseases in the world. In order to synthesize research for new developments, there is a demand for tools that can autonomously extract knowledge from biological literature.

One such tool is ScispaCy. ScispaCy is an open-source Python-based natural language processing (NLP) pipeline that is designed to analyze biomedical and scientific literature. This package contains the general Python-based spaCy models that use advanced NLP [24][37]. Along with the scispaCy library, there exists several pre-trained models for analyzing biological papers. Primarily, there is the pre-trained "*en_ner_bionlp13cg_md*" model which is a spaCy named entity recognition (NER) model trained on the BIONLP13CG corpus which is specific for cancer genetics [3]. NER is a type of NLP that identifies and categorizes named entities from text such as scientific literature. Named entities that are classified by the *en_ner_bionlp13cg_md* model include cancer, gene, tissue, cell, etc.

BERN2 is another NER model that uses a neural network and is specifically built for biomedical literature [45]. BERN2 was designed for any research paper that could be on Pubmed. BERN2 works by using cached annotations if possible, and if not uses a neural network NER model combined with rule based modelling to find entities. The classes of named entities BERN2 can extract include chemicals, disease, gene, and cell type or line.

Other tools exist for presenting extracted knowledge in new forms. KGen is a pipeline that designs knowledge graphs in a semi-automatic manner for unstructured biomedical literature that uses NLP and ontology linking [41]. KGen will graphically represent the knowledge extracted from the papers. A KGen graph, of the form $KG = (V, E)$, are generated by extracting Resource Description (RDF) triples from the text. An RDF triple, or semantic triple, represents a core relationship with a subject, predicate, and object (s, p, o), as is common in simple English language sentences. The graph is constructed by taking a set of triples, and adding them to the graphs where nodes are oriented in the form subject -¿ object. This yields the following graph:

$$E = (p_1, p_2, \ldots, p_n)$$

$$V = (s_1, s_2, \ldots, s_n, o_1, o_2, \ldots, o_n)$$

$$t_n = (s_n, p_n, o_n)$$

$$KG = (V, E) = (t_1, t_2, \ldots, t_n)$$

Ultimately, some of these subjects, objects, and predicates can be linked to an ontology. An ontology is defined as a domain of concepts, attributes, and relations. This gives an “Ontology-linked Knowledge graph”. Linking a knowledge graph to an established ontology can provide valuable context and background to a knowledge graph.

One of the major problems is that research done in breast cancer is highly technical and specific. General entity recognition models may not have the correct vocabulary or training necessary to recognize the entities that a research may be looking for.

This paper aims to evaluate current work and other related methods for autonomously analyzing biological papers, most specifically in how the results may contribute to the research of breast cancer subtyping. After evaluating the state of current tools, several contributions and improvements will be introduced that aim to improve process of creating custom NER models with specific entities, for specific research. These contributions will be evaluated on how well they extract knowledge from papers and the potential for application with regard to specific bio-informatics problems.

Chapter 2

BACKGROUND

As a function of the information revolution in the last few decades, there has also been a rapid increase in the availability of data in the bioinformatics space [28]. This has changed how medical research and treatment are administered. Especially since the mapping of the human genome, one of the highest areas of interest is that of genomic research. However, the methods with which we process the huge amounts of data are still evolving.

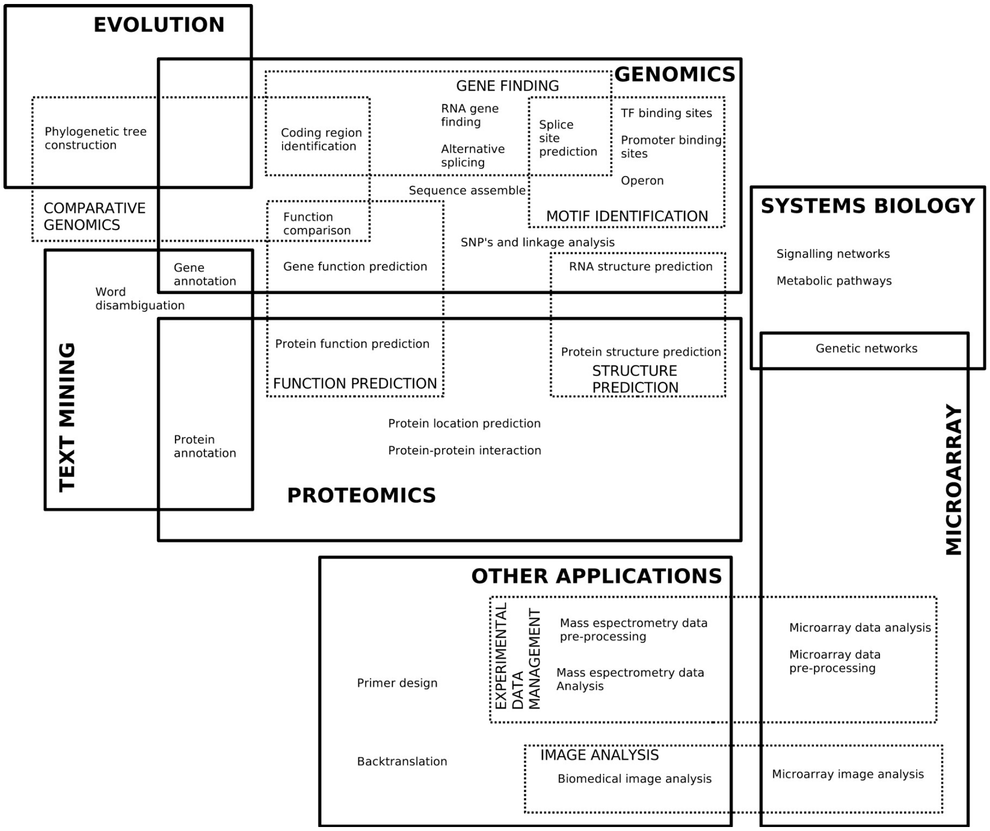


Figure 2.1: A breakdown of the bioinformatics field [28]

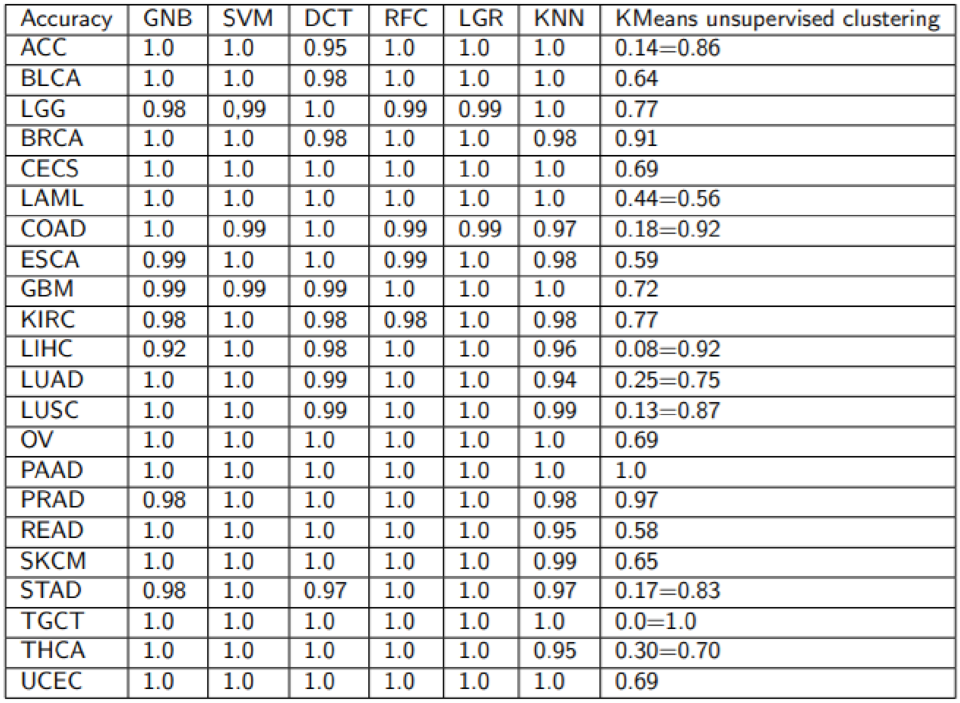
While not all inclusive, much of the work can be narrowed down into the following fields: “genomics, proteomics, microarrays, systems biology, evolution and text mining,” [28] and can be shown in the figure 2.1.

Genomics have been a part of breast cancer research for over 100 years. French surgeon Paul Broca first established that breast cancer may be hereditary in 1866, by analyzing the occurrence of 10 cases in 24 women of his wife’s family [48]. Further identification of the BRCA1 gene culminated in BRCA1 and then BRCA2, getting successfuly cloned in the 1990’s. It is estimated that up to 10% of breast cancer cases are hereditary, and with BRCA1, the onset of cancer appears two decades earlier [35]. The merging of 21st century computing power and the large amounts of medical research is one of the ways new developments can be made in breast cancer research and other medical fields.

Further work is necessary to fully understand and cure breast cancer. While breast cancer itself is widely studied, the relapse of breast cancer is not studied as well, despite being strongly associated with breast cancer mortality [1]. Furthermore, none of the three major American cancer registries record information on recurrence. Several approaches have been attempted to predict recurrence in breast cancer patients, including KDD, SEMMA and more. For feature selection, approaches varied between manual selection by medical experts, and others using feature selection algorithms. Some approaches focused largely on clinical data such as age, tumor size and stage, ER and HER2 status, and menopausal status. However, the accuracy varied widely. Common issues to this problem were lack of data, an imbalance of data, feature selection, interpret-ability, and an inability to evaluate success. Doctors are expected to explain diagnoses and medical decisions to patients, other doctors, and administrators. Artificial intelligence strategies can be difficult to explain, and if a physician cannot explain a diagnosis, then the method will never be excepted by the medical community[1].

2.1 Machine Learning in Bioinformatics

So far, numerous machine learning methods have been applied to cancer data, with promise in the classification results. Figure 2.2 shows a survey of core machine learning algorithms applied to a data-set for cancer classification[21]. The results show shockingly good accuracy, and unlock much of what will be discussed on this paper, they did not attempt to reduce dimmensionality, which may suggest over-fitting.



Accuracy	GNB	SVM	DCT	RFC	LGR	KNN	KMeans unsupervised clustering
ACC	1.0	1.0	0.95	1.0	1.0	1.0	0.14=0.86
BLCA	1.0	1.0	0.98	1.0	1.0	1.0	0.64
LGG	0.98	0,99	1.0	0.99	0.99	1.0	0.77
BRCA	1.0	1.0	0.98	1.0	1.0	0.98	0.91
CECS	1.0	1.0	1.0	1.0	1.0	1.0	0.69
LAML	1.0	1.0	1.0	1.0	1.0	1.0	0.44=0.56
COAD	1.0	0.99	1.0	0.99	0.99	0.97	0.18=0.92
ESCA	0.99	1.0	1.0	0.99	1.0	0.98	0.59
GBM	0.99	0.99	0.99	1.0	1.0	1.0	0.72
KIRC	0.98	1.0	0.98	0.98	1.0	0.98	0.77
LIHC	0.92	1.0	0.98	1.0	1.0	0.96	0.08=0.92
LUAD	1.0	1.0	0.99	1.0	1.0	0.94	0.25=0.75
LUSC	1.0	1.0	0.99	1.0	1.0	0.99	0.13=0.87
OV	1.0	1.0	1.0	1.0	1.0	1.0	0.69
PAAD	1.0	1.0	1.0	1.0	1.0	1.0	1.0
PRAD	0.98	1.0	1.0	1.0	1.0	0.98	0.97
READ	1.0	1.0	1.0	1.0	1.0	0.95	0.58
SKCM	1.0	1.0	1.0	1.0	1.0	0.99	0.65
STAD	0.98	1.0	0.97	1.0	1.0	0.97	0.17=0.83
TGCT	1.0	1.0	1.0	1.0	1.0	1.0	0.0=1.0
THCA	1.0	1.0	1.0	1.0	1.0	0.95	0.30=0.70
UCEC	1.0	1.0	1.0	1.0	1.0	1.0	0.69

Figure 2.2: A general survey of ML technique performance [21]

2.1.1 Feature Selection

When it comes to processing transcriptomic data, an ongoing issue is feature selection. With base pairs in a human in the millions, efficient algorithms that can find the most powerful of features are necessary for practical use of the vast amounts of information. One train of thought with feature selection is that most methods fall under 3 categories[10]. Filter methods focuses on the characteristics of individual genes and there relation to the target for classification. Filter methods tend to ignore more complicated relations but is relatively easy to use. Wrapper methods, generally considered to be more effective, is based around optimizing algorithm performance, by using an initial set of genes and augmenting it to learn the optimal number. The embedded method is similar to wrapper methods, but can combine multiple algorithms for feature selections[10].

One such way that feature selection can be done is a stacked auto encoder. An auto encoder is an unsupervised neural network the works at dimension reduction by deconstructing data and attempting to reassemble it, therefore identifying the minimal set of features necessary to do so. A stacked auto encoder works similarly, but by using multiple encoders in sequence. This helps improve dimension reduction for data sets with highly complex relationships.

One direct application of a stacked auto encoder was an attempt to predict cancer subtypes utilizing both genomic data as well as alternative splicing (AS) data. Alternative splicing is when alternate combinations of exons are joined together in the final stage of transcription, allowing multiple messenger RNA to be produced from a single gene. Studies have shown AS to be correlated with cancer development in the human body, which means that it has potential to be a strong predictor of cancer sub-types. This study used a stacked auto encoder on both gene and alternative splicing data for feature selection, before using an auto encoder on the combination of the two in order to predict cancer sub-type. This approach is illustrated in 2.3.

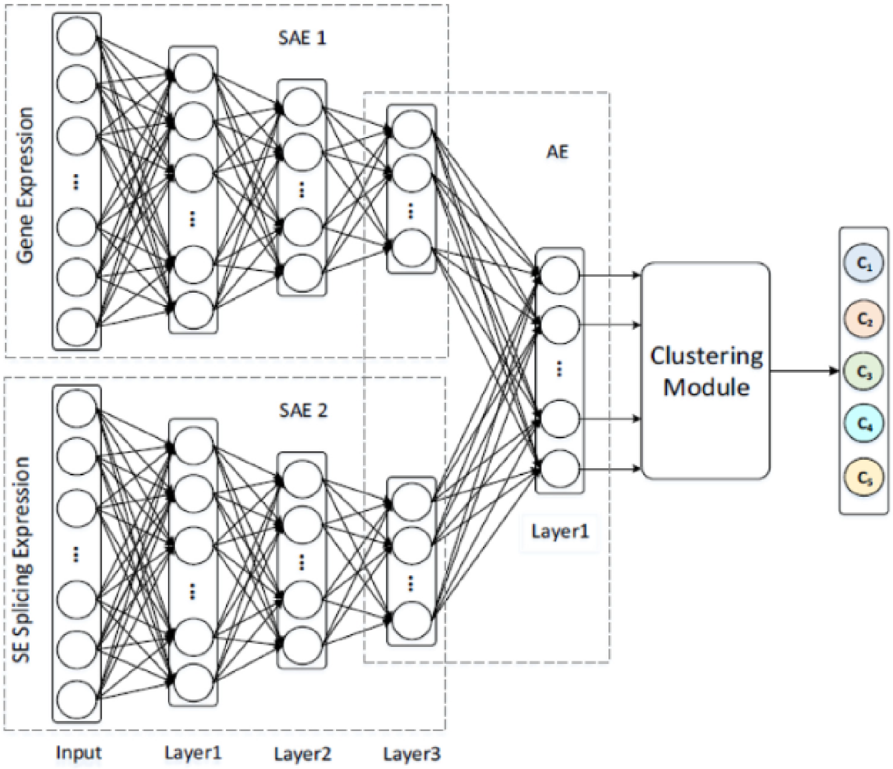


Figure 2.3: Stacked Auto Encoder Model [16]

This method outperformed PCA in terms of feature selection.

Another proposed method of feature selection is particle swarm optimization (PSO)[10]. In PSO, a group of candidate solutions referred to as particles navigate the search space by taking into account both local optimal position as well as global optimal position. One study combined this with C4.5 algorithm to perform gene selection[10]. In this method, the particle represents a subset of genes, where locations can be used to represent performance of the subset there. C4.5 is a decision tree algorithm known for efficiency in classification. C4.5 is used to evaluate performance of the candidate solution. This method reached 97% accuracy when applied to clinical studies. Another study combined PSO with synthetic minority oversampling technique to predict 5 year survivability[47]. Synthetic minority oversampling technique (SMOTE) is an approach for working with imbalanced data by created synthetic instances of the minority classification, in order to reduce the imbalance in the data. The new instances are created using a k-nearest-neighbors approach. In cancer data, many more patients today survive than do not, which creates a data imbalance. This results in low accuracy when it comes to predicting which patients will not survive. The study used several classifiers to compare accuracy of their model. There was mixed results, but the C5 classifier showed strong improvement using the hybrid model.

Another model combined minimum redundancy maximum relevance (MRMR) and cuckoo optimization algorithm (COA-HS) in two stages to narrow down the gene pool[14]. MRMR is a filter method that tries to select genes with the highest correlation to the target (maximum relevance) but with the least correlation to each other (minimum redundancy). COA is a method in which a ”nests” are analyzed for there probability of an ”egg” surviving in it. In this insance an egg represents a potential solution, and overtime, only the optimal solutions will remain in the nests. Harmony search (HS) is a musically inspired optimization method utilized as well. The results from MRMR are fed to a hybrid COA-HS search, and the results outperformed other methods for datasets on several types of cancer, including leukemia, prostate, lymphoma and colon cancer.

Automated methods reduce friction, but in these cases, fail to factor in valuable expert knowledge and current research, in the next section, the process of implementing expert knowledge will be discussed.

2.1.2 Knowledge Extraction

With the problem of feature selection, one factor that may improve selection is use of the vast amounts of medical research that has already been done in the relevant fields. As mentioned above, using medical expert or general prior knowledge is one of several ways to work on feature selection. One study established that injecting knowledge into machine learning models increased accuracy of cancer sub-type classification.

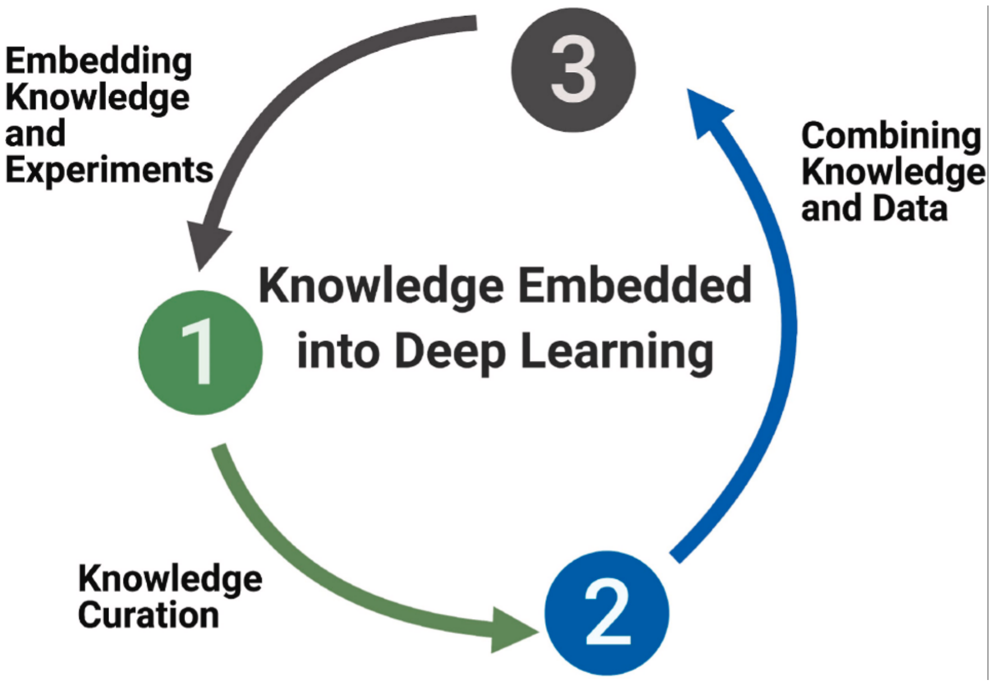

Embedding Knowledge and Experiments
3
Combining Knowledge and Data
1
Knowledge Embedded into Deep Learning
Knowledge Curation
2

Figure 2.4: A model of the cycle of knowledge extraction. [5]

By injecting knowledge from another paper, which identified three genes of interest (ESR1, ERBB2/HER2, AURKA) for identifying four different cancer sub-types (basal-like, HER2-enriched, luminal A, and luminal B). They then created a graph using k-nearest-neighbors where patients share an edge if they are of the same sub-type and are k-nearest-neighbors by distance using those three genes. By combining this knowledge graph with output from a deep learning neural network, the accuracy of the model increased from 75% to 80% [5].

This process of identifying valuable information in literature and applying it to bioinformatics models is a process that can be improved on. A primary motivation of this paper is to improve on this process by removing friction from the knowledge extraction process, and make for more effective extraction of relevant information from biological papers.

2.1.3 Knowledge Graphs

These results introduce another part of the puzzle for potential research. Knowledge graphs, meaning using a graph based data model has numerous upsides. A graph based architecture can illustrate complex relations, while maintaining flexibility[20]. In the field of bioinformatics, this model has several specific benefits. Knowledge graphs flexibility not only makes them more accepting of the rapidly growing and changing field of bioinformatics, but also makes it easier to merge the existing data, which is represented in numerous schema[18]. Graphs make it possible to represent ontologies and vocabularies, as well as executing complex queries with little computational expense[18]. A proposed use of a knowledge graph model is described by one paper as a digital library framework. This model would serve the ”needs of users (societies), provide information services (scenarios), organize information in usable ways (structure), present information in useful ways (spaces), and communicate information with users (streams)”[18]. This structure has several layers. The data layer may represent numerous formats covering several different knowledge bases. The processing layer performs preprocessing necessary for the data’s eventual merging. This yields the third layer, the knowledge graph. The graph itself has data organization nodes, and instances of those nodes. The fourth and fifth layer represent applications and there users that could be built on this knowledge graph. This paper created a prototype with a couple datasets, but introduced potential future work in introducing additional datasets.

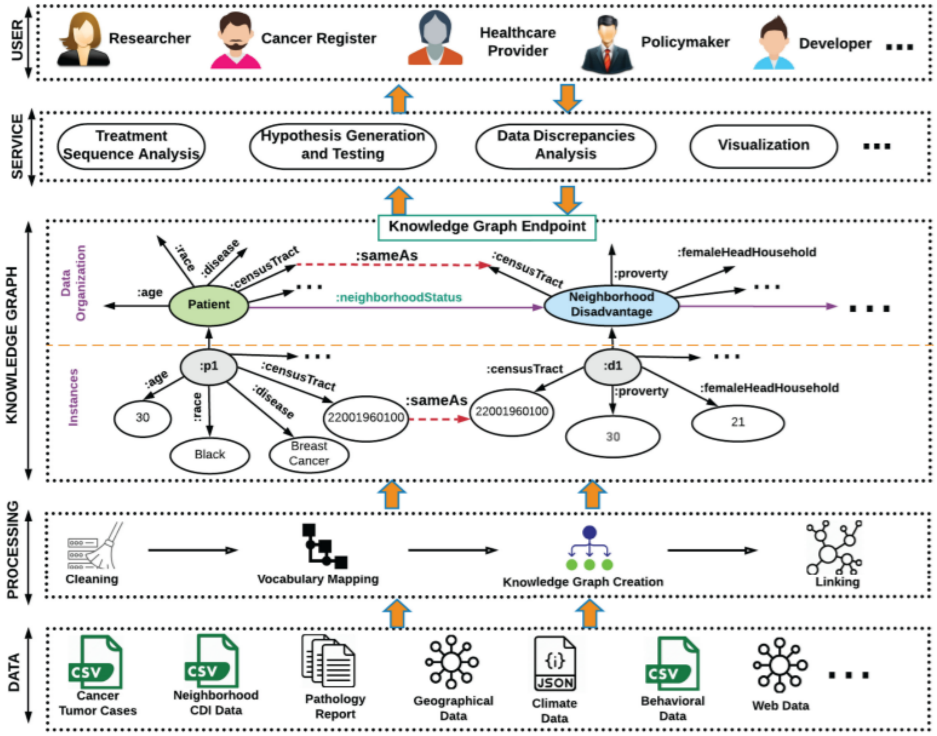


Figure 2.5: Knowledge Graph Model [16]

Knowledge graphs have also been used for predicting clinical outcomes. The paper used a knowledge graph to synthesis knowledge from several data sources, including CNA, methylation, gene expression, and miRNA. It used this information to try and predict three clinical outcomes[25]. Early or late stage, low or high grade, and short-term vs long-term survival (with 3 years being the benchmark). Edges in the graph were representative of similarities for two patients who have similar stats with regard to the above data sources. The study ultimately showed improvements in accuracy when integrating assorted data sources together.

The work in extracting knowledge to a knowledge graph may also include using natural language processing to analyze existing medical literature. Here, the knowledge graph utilizes semantic representations of medical knowledge. This research has great potential in making unseen connections between multiple sources of information. The SemRep system is a rule based program that tags medical entities from text from PubMed. It uses natural language processing methods such as part of speech tagging and named entity recognition. The paper furthermore incorporates quantitative metrics to identify the strength of relations found between nodes on the knowledge graph[11]. Disease specific information is used to identify more specific relationships.The overall method involved extracting literature on a subject and from that extracting the semantic associations. Then they normalized the information and injected the concept-specific embeddings. Then built the knowledge graph using the information, that came out as a representation of related concepts. Figure 2.6 show the model for this method[11].

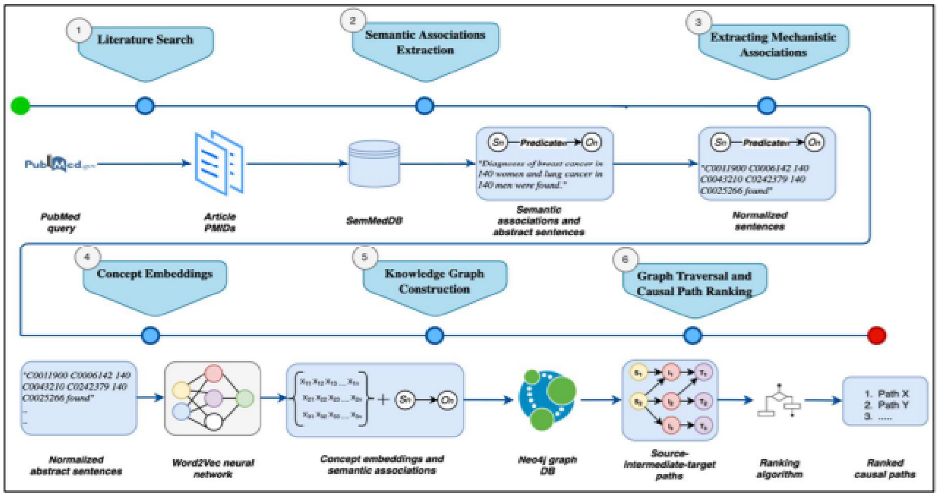

1
Literature Search
2
Semantic Associations Extraction
3
Extracting Mechanistic Associations
PubMed.gov
PubMed query
Article PMIDs
SemMedDB
Sn
Predicate
On
"Diagnoses of breast cancer in 140 women and lung cancer in 140 men were found."
Semantic associations and abstract sentences
"C0011900 C0006142 140 C0043210 C0242379 140 C0025266 found"
Normalized sentences
4
Concept Embeddings
5
Knowledge Graph Construction
6
Graph Traversal and Causal Path Ranking
"C0011900 C0006142 140 C0043210 C0242379 140 C0025266 found"
Normalized abstract sentences
Word2Vec neural network
+
Sn
On
Concept embeddings and semantic associations
Neo4j graph DB
Source-intermediate-target paths
Ranking algorithm
1. Path X
2. Path Y
3. .....
Ranked causal paths

Figure 2.6: Knowledge Graph utilizing NLP techniques to extract knowledge [16] [11]

Overall knowledge graphs can be a valuable way to both visualize extracted knowledge as well as a valuable data science tool for applying extracted knowledge. An improved entity recognition model could create improved knowledge graphs that have great value to researchers.

2.1.4 Other Methods

Another study sought to use machine learning to analyze the relation of the combined information of miRNA and lncRNA to breast cancer and neoplasm scenarios [17]. A neoplasm is an abnormal growth similar to cancer. The combine the data sets, the lncRNA (which is the longer of the two) was sequence aligned to the miRNA. Feature selection was done using k-mers and energy folding values, among others. One class SVM was then used to identify outliers, whose output was then fed to supervised decision tree and SVM models.

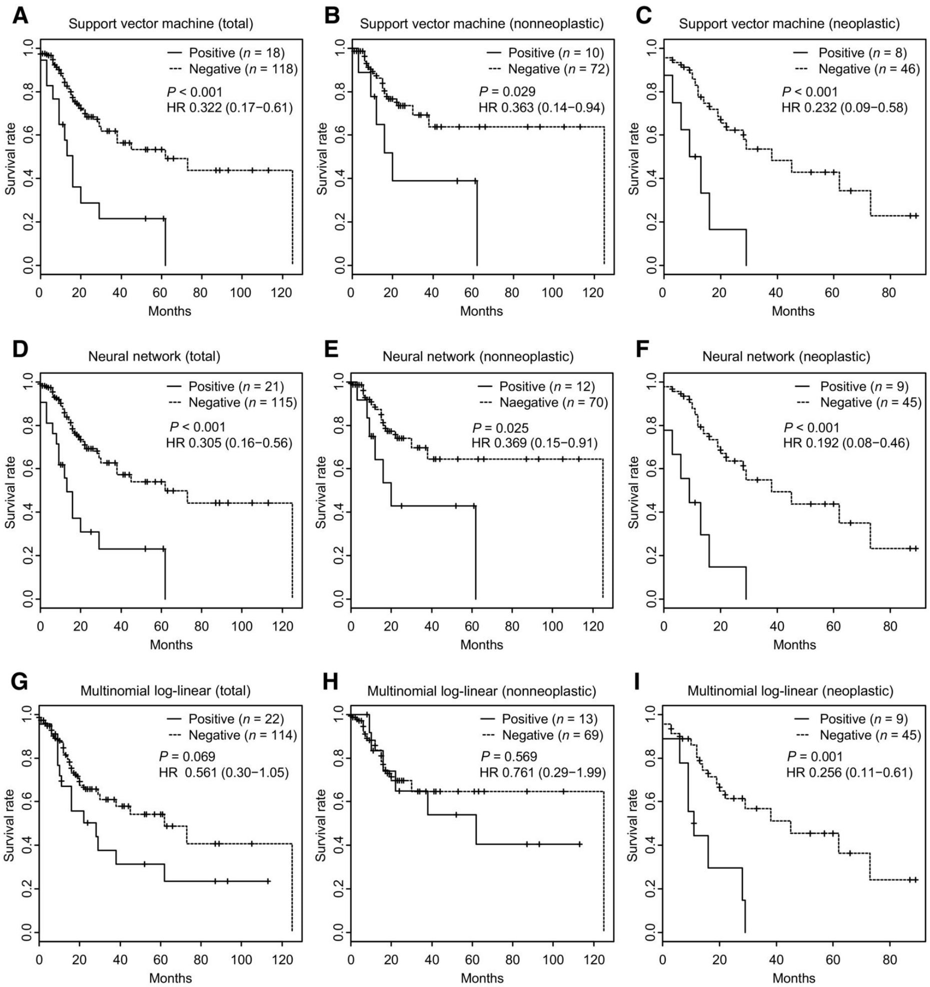


Figure 2.7: SVM Performance Analysis[52]

This approach achieved extremely good accuracy, which verifies the potential of combining some of our new bioinformatics data to yield improvements in model performance.

For pancreatic cancer, 3 genes of interest are MUC1, MUC2, and MUC4[52]. Several models have been attempted to classify prognosis of pancreatic cancer using these genes. A classic SVM classifier and two varieties of neural networks. Support vector machine (SVM) models are based on creating a plane or several planes of varying dimmensionality to try and graphically separate different labels. Neural nets were used from the nnet package, with a further model being a neural net that utilizes the multinomial log linear model using the multinom function, which is an expression for regression models. Both models had good results with regard to mapping prognosis as survival rate after a variable amount of months. This study included both neoplastic and nonneoplastic samples. A cell is neoplastic is if it has been transformed beyond its ability to perform normal body processes. The results for this study can be seen above.

Ultimately the methods above fail to utilize the vast body of research available regarding breast cancer, except by largely manual efforts. This paper will look to create a method for which researchers may use in order to get better results in the bioinformatic study of breast cancer and other diseases.

Chapter 3

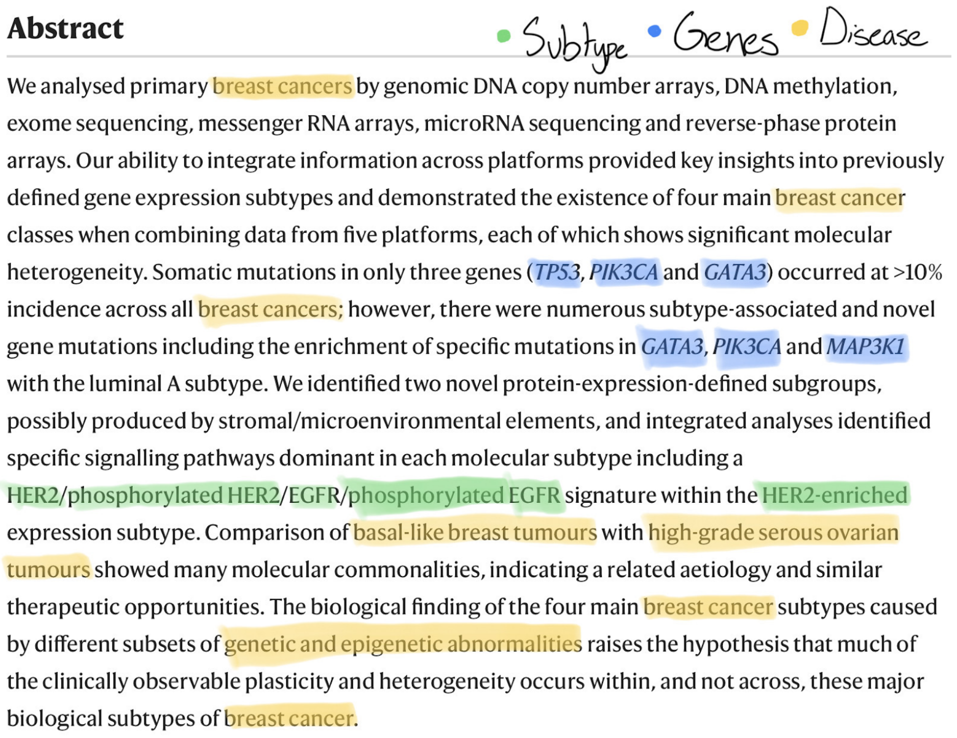
METHODS

3.1 Comparitive Analysis

The performance of several tools, namely ScispaCy, BERN2 and KGEN were investigated for their ability to extract information relative to the breast cancer subtyping problem. We initially selected two biological questions relating to breast cancer research and aimed to see how well these tools extract information relative to these questions from breast cancer subtyping papers.

1. Blows *et al.* found that low TP53 mutation frequency in luminal A (12%) and a higher frequency in luminal B (29%) cancers [9]. Have other researchers found the interesting change in mutation rate in TP53 between luminal A and luminal B?
2. Koboldt *et al.* found a luminal expression signature of ESR1, GATA3, FOXA1, XBP1 and MYB [26]. Have other researchers found similar signatures?

This approach was modelled to simulate how these tools might be used by biologists seeking information specific to a particular field of research. The papers used to test these methods on were chosen by biologist, on the basis of there relevance to breast cancer, primarily those that dealt with genomics in breast cancer subtyping. That is to say, the papers had immediate relevance to the questions above.The papers were annotated to provide a baseline to compare the NER efforts of ScispaCy and BERN2 with an example of which is shown in figure 3.1.



Abstract

We analysed primary breast cancers by genomic DNA copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays. Our ability to integrate information across platforms provided key insights into previously defined gene expression subtypes and demonstrated the existence of four main breast cancer classes when combining data from five platforms, each of which shows significant molecular heterogeneity. Somatic mutations in only three genes (*TP53*, *PIK3CA* and *GATA3*) occurred at >10% incidence across all breast cancers; however, there were numerous subtype-associated and novel gene mutations including the enrichment of specific mutations in *GATA3*, *PIK3CA* and *MAP3K1* with the luminal A subtype. We identified two novel protein-expression-defined subgroups, possibly produced by stromal/microenvironmental elements, and integrated analyses identified specific signalling pathways dominant in each molecular subtype including a HER2/phosphorylated HER2/EGFR/phosphorylated EGFR signature within the HER2-enriched expression subtype. Comparison of basal-like breast tumours with high-grade serous ovarian tumours showed many molecular commonalities, indicating a related aetiology and similar therapeutic opportunities. The biological finding of the four main breast cancer subtypes caused by different subsets of genetic and epigenetic abnormalities raises the hypothesis that much of the clinically observable plasticity and heterogeneity occurs within, and not across, these major biological subtypes of breast cancer.

Figure 3.1: An example of a manually annotated paper [26]

3.2 Annotations For Specific Problems

Reliance on large, established data sets reduces the capability of existing NLP tools to be applied to more specific problems. The creation of these corpus are at the same time too tall a task to be done for every niche topic using existing tools. To train a spaCy pipeline, sets of annotations are necessary. An annotation looks like the following:

(”We identified two novel protein-expression-defined subgroups, possibly produced by stromal/microenvironmental elements, and integrated analyses identified specific signalling pathways dominant in each molecular subtype including a HER2/phosphorylated HER2/EGFR/phosphorylated EGFR signature within the

HER2-enriched expression subtype.”, ”entities”:
[(24,60,”Subtype”),(231,250,”Subtype”),(251,275,”Subtype”),(276,334,”Subtype”)])

The use of string indices to pinpoint the entities makes the creation of these annotations tedious if done manually. When factoring in the desire for a large number of samples for a training set of data, this can be an extremely laborious task.

Our initial work is focused on improvement on an existing tool, the helps to expedite this process. The spaCy annotation tool from agate team begins with several good features[2].

- Custom labels for entities
- Assign labels by highlighting (calculates indices automatically)
- Formats annotations automatically

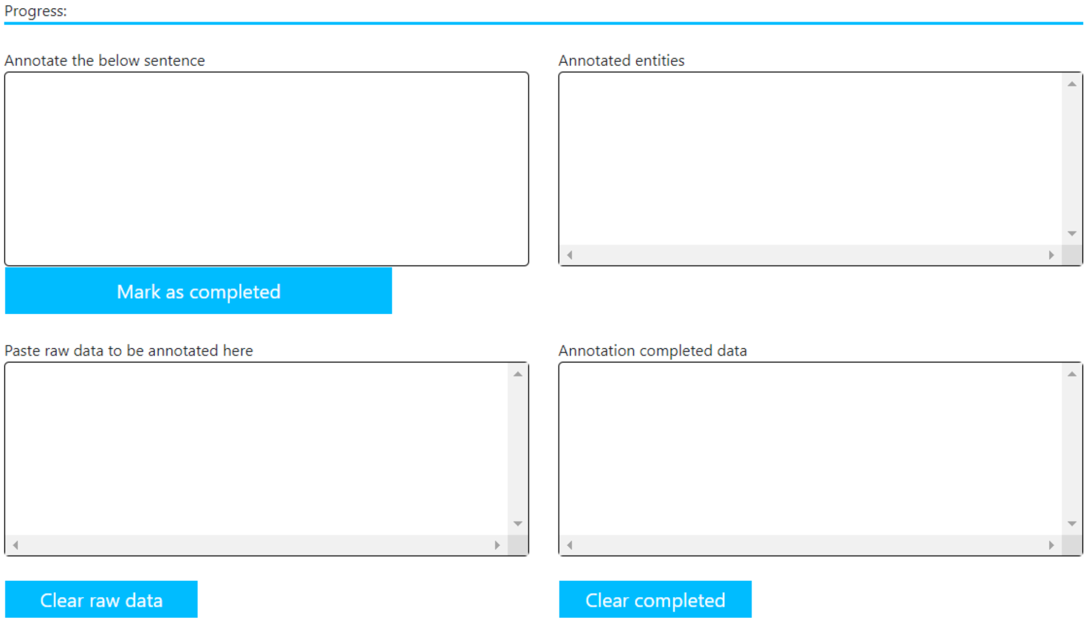
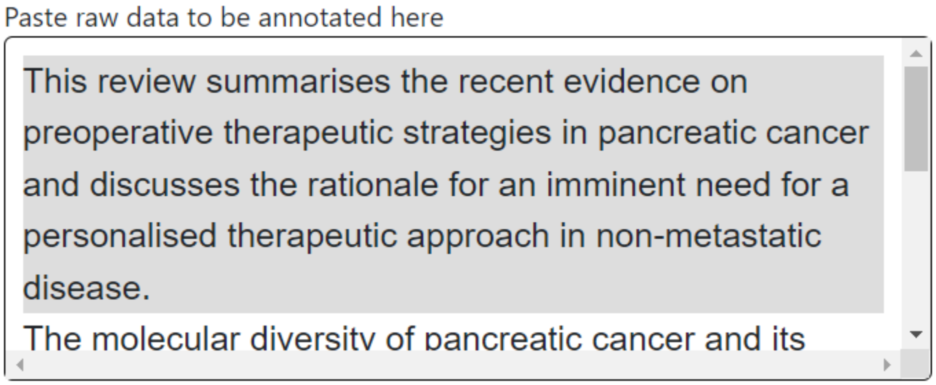

Progress:
Annotate the below sentence
Annotated entities
Mark as completed
Paste raw data to be annotated here
Annotation completed data
Clear raw data
Clear completed

Figure 3.2: The Annotation Tool [2]

These are notable improvements on doing the annotations manually. Assigning labels by highlighting is the primary feature, as it makes assigning names to entities easy. However, there are also drawbacks. The only way to input text is to paste text into the tool. Furthermore, the text needs to be formatted with new lines for each sentence, requiring manual preprocessing. Likewise, the only way to get the data out is to copy and paste out the completed annotations. In addition, the labels start at a clean slate every load of the tool, which makes inconsistency a risk, especially when doing collaborative research.

The new version of the annotation tool contains several new features. The focus of these features is to removing friction from this process. The software can automatically preload an abstract into the annotation tool using URL parameters for any pubmed ID for which the abstract is publicly available. This performs the necessary preprocessing of the text in order to prepare it for annotation. For example, the URL [host]/spacyannotation/?pubmedID=33248227 preloads a paper into the tool with the text processed to prepare it for annotation.



This review summarises the recent evidence on preoperative therapeutic strategies in pancreatic cancer and discusses the rationale for an imminent need for a personalised therapeutic approach in non-metastatic disease.

The molecular diversity of pancreatic cancer and its

Figure 3.3: The input text box from the annotation tool[19]

This is the abstract from the paper *Reshaping preoperative treatment of pancreatic cancer in the era of precision medicine.* 33248227 is the pubmed ID for this paper.

For any paper whose abstract is publically available, this URL parameter will load the abstract into the tool for annotation.

Another method was used for developing the test corpus. A group of files can be preloaded into the tool and accessed using a different URL parameter. With a collection of texts, you can quickly cycle through annotating papers. Using python scipts, a folder with text files containing text can be prepared to be annotated. The URL [host]/spacyannotation/?paper=n will load these papers into the tool where n is an index from 1 to the number of papers in the folder. This method was used for creating the training data.

These methods have several advantages over copy and paste as input. First, this manages some of the prepossessing that might otherwise need to be done manually, most notably by separating the text by sentence. Secondly, the use of URL parameters makes for permanent URLs, meaning a link or links can be shared to collaborate on annotating text. As volume is relevant when building the training data, being able to collaborate is critical. The tool also can be adjusted to have preset labels, which adds consistency to a collaborative effort. Having all partners using the exact same labels adds to the cohesiveness of the resulting dataset.

For both of these methods, the tool will save the results of sessions in order to build a single group of annotations. Another addition we have is the ability to save the formatted annotations as a .txt file, instead of copying and pasting the results.

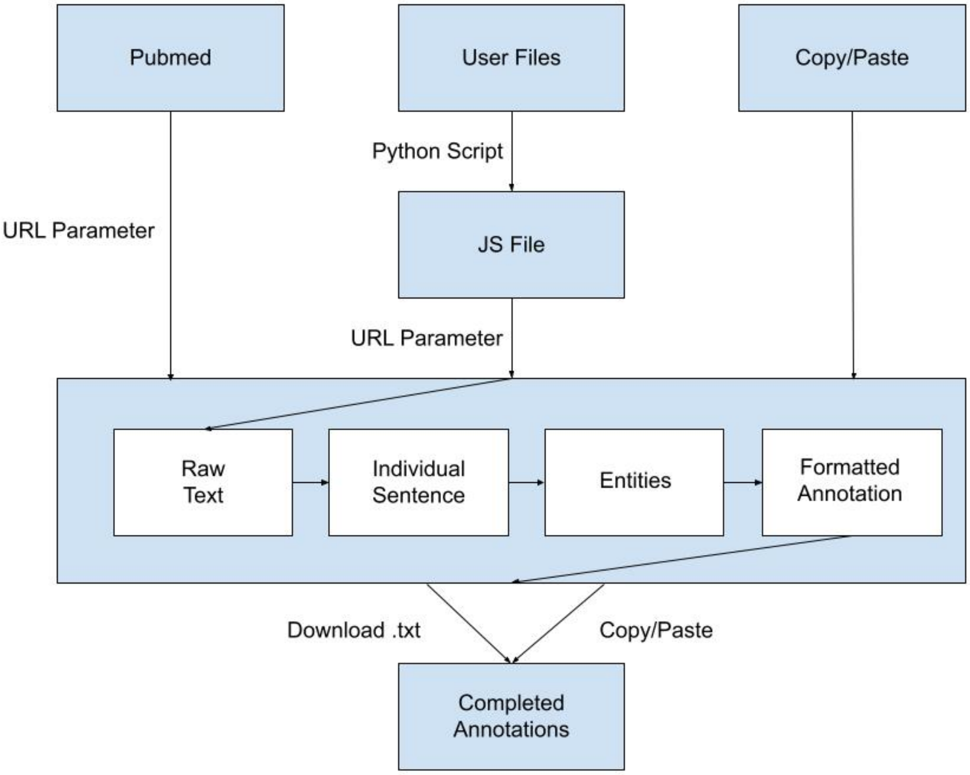
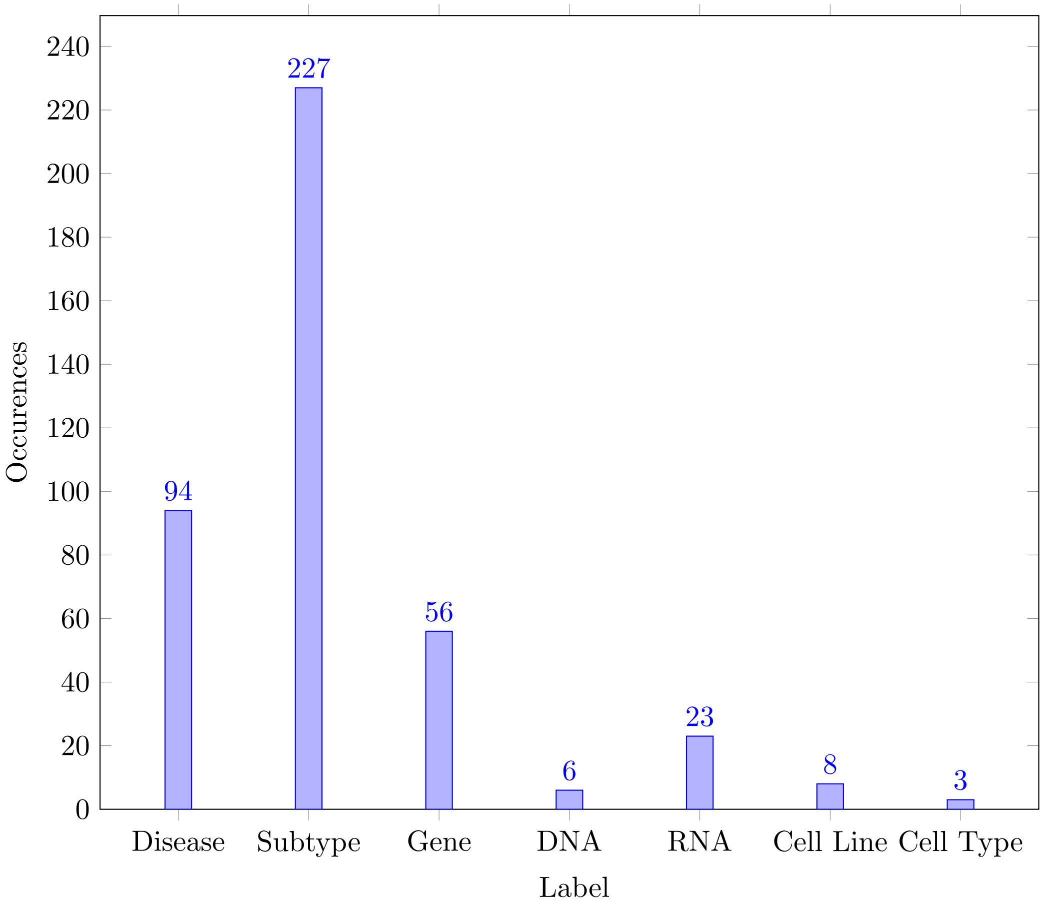


Figure 3.4: Annotation Tool Flow Chart

The end goal is to make creating a data set quickly and cooperatively in order to train named entity recognition pipelines for specific uses. For the training corpus, we created annotations for the abstracts of 25 papers, representing over 250 sentences. The papers were selected by members of the Bioinformatics Research Group (BIRG) for relation to breast cancer subtyping.



3.3 Custom NER Performance Analysis

For the purposes of comparison, in one run we chose the same labels as another NER model, BERN2 [45]. Advanced Biomedical Entity Recognition and Normalization (BERN2). BERN2 utilizes a multi task neural network model, and showed good results in identifying 9 types of entities.

- Gene
- Disease
- Chemical
- Species
- Mutation
- Cell Line
- Cell Type
- DNA

- RNA

BERN2 used several models to identify entities of these types with fairly good results. It attempted both a neural network as well as a rules based approach, and ultimately utilizes a hybrid approach that can identify entities up over 90% at a time. The aforementioned labels are also a very appropriate for our topic, breast cancer subtyping. The papers selected by BIRG were annotated by several members for these labels.

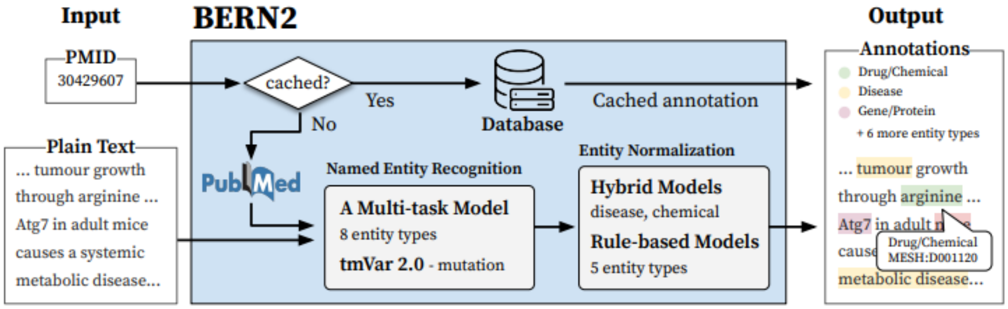


Figure 3.5: BERN2 Flow Chart[45]

To compare performance, we discussed both the overall accuracy and the accuracy with regard to each label individually. This will provide a comprehensive look at the comparative performance of our entity recognition model. In addition we calculated precision, recall, and f1 score. These values are calculated from the classification measures of true positive, false positive, false negative, true negative.

- True Positive (TP): Predicted Positive and was Positive
- False Positive (FP): Predicted Positive but was Negative
- False Negative (FN): Predicted Negative but was Positive
- True Negative (TN): Predicted Negative and was Negative

The values above, once accounted for in the results derived from the experiment, can be used to calculate the following.

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$
$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

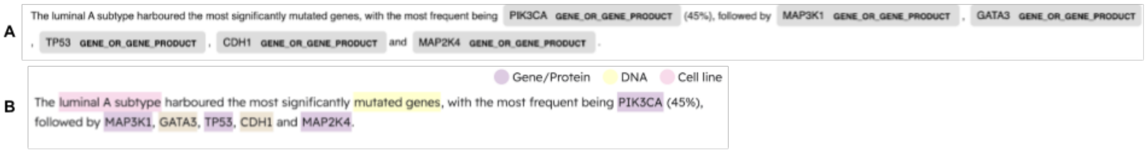
Chapter 4

RESULTS

4.1 Comparitive Analysis

In the initial investigation we manually compared two leading entity extraction methods: ScispaCy and BERN2. The results of this analysis are shown in Figure 4.1. We observe that both methods correctly identify genes and that BERN2 incorrectly classifies subtype, while ScispaCy omits it. We also observe a misidentification of gata3 and cdh1, and finally, we note that ScispaCy does not identity luminal A.

Similar results are observed for Question 2. The sentences that originates this question is ”One of the most dominant features is high mRNA and protein expression of the luminal expression signature, which contains ESR1, GATA3, FOXA1, XBP1 and MYB; the luminal/ER+ cluster also contained the largest number of significantly mutated genes.” For this sentence, our analysis showed that both ScispaCy and BERN2 identified the genes, BERN2 incorrectly classifies the subtype, ScispaCy omits the subtype, and ScispaCy misidentifies luminal and ER+ as genes.



The luminal A subtype harboured the most significantly mutated genes, with the most frequent being PIK3CA GENE_OR_GENE_PRODUCT (45%), followed by MAP3K1 GENE_OR_GENE_PRODUCT , GATA3 GENE_OR_GENE_PRODUCT , TP53 GENE_OR_GENE_PRODUCT , CDH1 GENE_OR_GENE_PRODUCT and MAP2K4 GENE_OR_GENE_PRODUCT .

The luminal A subtype harboured the most significantly mutated genes, with the most frequent being PIK3CA (45%), followed by MAP3K1, GATA3, TP53, CDH1 and MAP2K4.

Figure 4.1: Named entity extraction results for Question 2 shown for (A) ScispaCy and (B) BERN2.

4.2 Annotations For Specific Problems

The tool can ultimately be used by utilizing npm live-server [https://github.com/GrifH/spacyannota Running the command live-server in the directory of the tool will launch the tool as a locally hosted webpage. From, here the url http://127.0.0.1:8080/?pubmedID=33248227 automatically loads the abstract from the paper Reshaping preoperative treatment of pancreatic cancer in the era of precision medicine. From here annotating is very easy.

Table 4.1: Annotation Tool Comparison[2][19]

Features	Agate-Team Tool	New Tool
Copy/Paste Input	✓	✓
Pubmed ID Input		✓
File Load Input		✓
Permanent URL		✓
Permanent Labels		✓
Copy/Paste Output	✓	✓
File Download Output		✓
File Load Input		✓

Biologists using the tool have reported that the tool has greatly reduced the time it takes to annotate biological papers, especially with the pre-load tool from the command line. While a large magnitude of papers still need to be researched for annotating, this tool makes the annotating process itself more efficient. The permanent urls make annotating easier in terms of distributing and sharing work.

For analyzing these results, it is difficult to pin down a baseline, especially in the early stage of this research. Future work on this should focus on both subjective and objective analysis of the effectiveness of this tool for creating annotations, as well as the utility of those annotations. If they can be used to create specific use data sets for natural language processing applications, then the utility of this tool is validated to a higher degree.

4.3 Custom NER Performance Analysis

For the actual performance of the annotations the overall recognition was analyzed. This was measured by looking at if an entity was annotated, was any entity recognized by spaCy or BERN2. This would not punish an incorrect labelling if a label was called for. In this matter, initial results yielded approximately 65% overall accuracy for the custom spaCy NER model. Compared to the 80% BERN2 accuracy, our result was notably worse. However, given the relatively small sample size represented by the corpus, this is an encouraging result. The true gain relative to our goals come when looking at individual labels, as will be observed below.

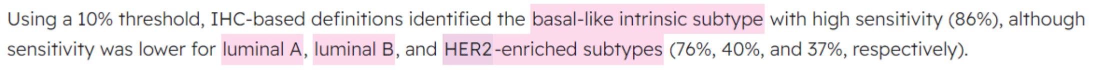
The tables below describe results from our retrained spaCy Model and that of BERN2. These labels include both ones common to BERN2 but also ones unique to our problem, subtype and gene product, treatments and tests, although not all of these labels occurred very often. The most common, as one might imagine, were disease, subtype and gene. This is because most of the papers selected were focused on these topics, and less so necessarily on treatments, symptoms or other cancer adjacent topics.

First, table 4.2 performance from BERN2 on our testing set.

Table 4.2: BERN2 Results[45]

Class	Precision	Recall	F1	Accuracy
Disease	0.96	0.47	0.63	0.46
Gene	0.67	0.85	0.74	0.59
Chemical	0.25	1	0.4	0.25
DNA	0.42	1	0.6	0.43
RNA	1	1	1	1
Cell Type	0.8	1	0.89	0.8
Cell Line	0.09	1	0.16	0.09
Mutation	N/A	N/A	N/A	N/A
Species	1	1	1.0	1
Overall	**0.68**	**0.69**	**0.68**	**0.52**

Looking at this table, the results seem very poor, especially relative to the quoted 80% from above. This was because of subtypes, an element critical to our specific problem, were most often classified incorrectly as cell line's or genes as opposed to a disease, which is where they hypothetically should be within the constraints of these labels. Take the following as an example:



Using a 10% threshold, IHC-based definitions identified the basal-like intrinsic subtype with high sensitivity (86%), although sensitivity was lower for luminal A, luminal B, and HER2-enriched subtypes (76%, 40%, and 37%, respectively).

Figure 4.2: BERN2 Example[45]

For any number of reasons, subtypes were often classified incorrectly. In this example, this list of subtypes, clearly prefaced as subtypes in the text, are miss labelled as cell lines. This is not only generally incorrect, but specifically unproductive within the framework of our specific research into breast cancer subtypes.

Our own method produced the following results for some common labels and some of our own:

Table 4.3: Custom NER Results

Class	Precision	Recall	F1	Accuracy
Disease	0.89	0.73	0.8	0.67
Subtype	0.88	0.73	0.8	0.67
Gene	0.81	0.65	0.72	0.56
Gene Product	0.33	0.083	0.13	0.07
DNA	0	0	0	0
RNA	1	1	1.0	1
Cell Type	0	0	0	1
Cell Line*	N/A	N/A	N/A	N/A
Mutation*	N/A	N/A	N/A	N/A
Species*	N/A	N/A	N/A	N/A
Chemical*	N/A	N/A	N/A	N/A
Treatment	1	1	1.0	1
Receptor	0.83	1	0.91	0.83
Test	1	0.8	0.89	0.8
Overall	**0.86**	**0.66**	**0.37**	**0.59**

* Our annotating team did not annotate any labels of these types, so performance data on these is non-applicable.

While the accuracy ratings in this table are underwhelming, a positive result is that it learned unique labels such as subtype, and established good precision and some degree of accuracy. Compared to the BERN2 results, this is about as accurate, and the performance was superior in the categories most relevant to use, specifically, diseases and subtypes. Genes were recognized with similar accuracy, although gene product (proteins) struggled considerably. Given that BERN2 most often mislabelled subtypes as diseases and genes, being able to recognize subtypes without damaging performance to other diseases and genes is a very good result.

Chapter 5

CONCLUSIONS AND DISCUSSION

Experiments with existing tools yielded lackluster results. The entity recognition was not able to clue in entities that were particular to the breast cancer subtyping problem. As a first step towards improving on this problem, one would have to create a new dataset. Spacy, and by extension KGen, are trained on an existing corpus that have been annotated to show spacy what entities and labels to look for. Making these annotations is a tedious and slow process, and numerous papers must be annotated for any degree of training to be effective. To this end, we have improved an existing annotation tool to improve scalability, and create a tool that can be used to quickly annotate papers and build new datasets and evaluated the effectiveness of these custom annotations[2]. These experiments have established the possibility of tackling specific problems by novel named entity recognition models. While larger corpus will long have an advantage in broad topics, the models ability to pick up labels from just the training data with some degree of accuracy shows this method can be applied to specific topics, something that can be of great use for researchers.

Research being done at higher institutions is often technical enough that broad topic recognition may not be as valuable as a specific search for named entities. The overhead of making the annotations and training the model are largely trivial in the scope of long term research, and with some future work, may make significant contributions to medical research.

The problem can be summarized as two problems, the search for information and the synthesis of it. The search for relevant information is still largely done manually, as it was for the contributions of this paper. The contribution of this paper will ultimately be a larger part of the latter problem, synthesis. Previous applications in bioinformatics have been discussed in this paper already, but one that was focused on was KGen.

KGen, a knowledge graph generator, generates triples of subject, predicate, and object to create a graph, with subject and object being vertices, and the predicate as an edge[41]. KGen can integrate components from spaCy pipelines. KGen uses ScispaCy as its default NER models. KGen attempts to use ScispaCy to link entities to an ontology, a database of biomedical information. The potential for well-founded knowledge graphs makes integration with the work in this paper a natural next step. Linking entities to each other and to important background for the entities is a valuable method of synthesizing critical information together. Our custom model can be added to KGen and function in terms of graph creation. However it does not create any ontological links when utilizing the custom NER model. The default model, unfortunately yeilded similar results for our breast cancer subtyping. KGen graphs also tend to be ineffective at linking biomedical entities, as NER is only a part of their pipeline for ontology linking. Their own paper reported good results for ontology linking. Projects like KGen, in combination with more specific NER work like this paper, can be a part of both problems above, search and synthesis. Effective graph creation taking NER into consideration, can be an effective way of extracting information from papers, and effective ontology creation can provide depth and understanding for these graphs.

The experiments here had several limitations. The annotations used to create this NER model were created by students in the BIRG research group, there was inevitably a degree of human error. This is also a factor when evaluating the performance of the models above. The process of loading papers in for annotation is also currently used only for PubMed and only for abstracts that are publically available. There are other databases that could be integrated for a more robust experience.

BIBLIOGRAPHY

[1] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva. Predicting breast cancer recurrence using machine learning techniques: A systematic review. *ACM Comput. Surv.*, 49(3), oct 2016.

[2] Agate-Team. spacyannotation, 2020.

[3] AllenAI. Scispacy, 2022.

[4] E. H. Allott, S. M. Cohen, J. Geradts, X. Sun, T. Khoury, W. Bshara, G. R. Zirpoli, C. R. Miller, H. Hwang, L. B. Thorne, et al. Performance of three-biomarker immunohistochemistry for intrinsic breast cancer subtyping in the amber consortium. *Cancer Epidemiology and Prevention Biomarkers*, 25(3):470–478, 2016.

[5] P. Anderson, R. Gadgil, W. A. Johnson, E. Schwab, and J. M. Davidson. Reducing variability of breast cancer subtype predictors by grounding deep learning models in prior knowledge. *Computers in Biology and Medicine*, 138:104850, 2021.

[6] H. Azim, S. Michiels, F. Zagouri, S. Delaloge, M. Filipits, M. Namer, P. Neven, W. Symmans, A. Thompson, F. André, S. Loi, and C. Swanton. Utility of prognostic genomic tests in breast cancer practice: The impakt 2012 working group consensus statement. *Annals of oncology : official journal of the European Society for Medical Oncology*, 24(3):647—654, March 2013.

[7] F. B. Azzouz, B. Michel, H. Lasla, W. Gouraud, A.-F. François, F. Girka, T. Lecointre, C. Guérin-Charbonnel, P. P. Juin, M. Campone, et al. Development of an absolute assignment predictor for triple-negative breast

cancer subtyping using machine learning approaches. *Computers in Biology and Medicine*, 129:104171, 2021.

[8] R. R. Bastien, Á. Rodríguez-Lescure, M. T. Ebbert, A. Prat, B. Munárriz, L. Rowe, P. Miller, M. Ruiz-Borrego, D. Anderson, B. Lyons, et al. Pam50 breast cancer subtyping by rt-qpcr and concordance with standard clinical molecular markers. *BMC medical genomics*, 5(1):1–12, 2012.

[9] F. M. Blows, K. E. Driver, M. K. Schmidt, A. Broeks, F. E. Van Leeuwen, J. Wesseling, M. C. Cheang, K. Gelmon, T. O. Nielsen, C. Blomqvist, et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS medicine*, 7(5):e1000279, 2010.

[10] K.-H. Chen, K.-J. Wang, K.-M. Wang, and M.-A. Angelia. Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. *Applied Soft Computing*, 24:773–780, 2014.

[11] A. Daowd, M. Barrett, S. Abidi, and S. S. R. Abidi. A framework to build a causal knowledge graph for chronic diseases and cancers by discovering semantic associations from biomedical literature. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 13–22, 2021.

[12] J. J. de Ronde, J. Hannemann, H. Halfwerk, L. Mulder, M. E. Straver, M.-J. T. Vrancken Peeters, J. Wesseling, M. van de Vijver, L. F. Wessels, and S. Rodenhuis. Concordance of clinical and molecular breast cancer subtyping in the context of preoperative chemotherapy response. *Breast cancer research and treatment*, 119(1):119–126, 2010.

[13] X. Du, X.-Q. Li, L. Li, Y.-Y. Xu, and Y.-M. Feng. The detection of esr1/pgr/erbb2 mrna levels by rt-qpcr: a better approach for subtyping breast cancer and predicting prognosis. *Breast cancer research and treatment*, 138(1):59–67, 2013.

[14] V. Elyasigomari, D. Lee, H. Screen, and M. Shaheed. Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. *Journal of Biomedical Informatics*, 67:11–20, 2017.

[15] C. M. Focke, P. J. van Diest, and T. Decker. St gallen 2015 subtyping of luminal breast cancers: impact of different ki67-based proliferation assessment methods. *Breast cancer research and treatment*, 159(2):257–263, 2016.

[16] Y. Guo, X. Shang, and Z. Li. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. *Neurocomputing*, 324:20–30, 2019. Deep Learning for Biological/Clinical Data.

[17] J. Gutiérrez-Cárdenas and Z. Wang. Classification of breast cancer and breast neoplasm scenarios based on machine learning and sequence features from lncrnas-mirnas-diseases associations. *Interdisciplinary sciences, computational life sciences*, 2021.

[18] S. Hasan, D. Rivera, X.-C. Wu, E. B. Durbin, J. B. Christian, and G. Tourassi. Knowledge graph-enabled cancer data analytics. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1952–1967, 2020.

[19] G. Hawblitzel. spacyannotation, 2022.

[20] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4), jul 2021.

[21] A. Hooshmand. Machine learning against cancer: Accurate diagnosis of cancer by machine learning classification of the whole genome sequencing data, 2020.

[22] C. Horr and S. A. Buechler. Breast cancer consensus subtypes: A system for subtyping breast cancer tumors based on gene expression. *NPJ breast cancer*, 7(1):1–13, 2021.

[23] S. Janeva, T. Z. Parris, S. Nasic, S. De Lara, K. Larsson, R. A. Audisio, R. Olofsson Bagge, and A. Kovács. Comparison of breast cancer surrogate subtyping using a closed-system rt-qpcr breast cancer assay and immunohistochemistry on 100 core needle biopsies with matching surgical specimens. *BMC cancer*, 21(1):1–10, 2021.

[24] S. JUGRAN, A. KUMAR, B. TYAGI, and V. ANAND. Extractive automatic text summarization using spacy in python & nlp. In *Extractive Automatic Text Summarization using SpaCy in Python & NLP*, pages 582–585, 03 2021.

[25] D. Kim, J.-G. Joung, K.-A. Sohn, H. Shin, Y. R. Park, M. D. Ritchie, and J. H. Kim. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. *Journal of the American Medical Informatics Association*, 22(1):109–120, 07 2014.

[26] D. Koboldt, R. Fulton, M. McLellan, H. Schmidt, J. Kalicki-Veizer, J. McMichael, L. Fulton, D. Dooling, L. Ding, E. Mardis, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[27] S. Lakis, V. Kotoula, A. G. Eleftheraki, A. Batistatou, M. Bobos, T. Koletsa, E. Timotheadou, S. Chrisafi, G. Pentheroudakis, A. Koutras, et al. The androgen receptor as a surrogate marker for molecular apocrine breast cancer subtyping. *The Breast*, 23(3):234–243, 2014.

[28] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 03 2006.

[29] E. Lips, L. Mulder, J. De Ronde, I. Mandjes, B. Koolen, L. Wessels, S. Rodenhuis, and J. Wesseling. Breast cancer subtyping by immunohistochemistry and histological grade outperforms breast cancer intrinsic subtypes in predicting neoadjuvant chemotherapy response. *Breast cancer research and treatment*, 140(1):63–71, 2013.

[30] M. Liu, Z. Wang, T. Tan, Z. Chen, X. Mou, X. Yu, Y. Deng, G. Lu, and N. He. An aptamer-based probe for molecular subtyping of breast cancer. *Theranostics*, 8(20):5772, 2018.

[31] Q. Liu, B. Cheng, Y. Jin, and P. Hu. Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data. *Journal of biomedical informatics*, 125:103958, 2022.

[32] F. Lüönd, S. Tiede, and G. Christofori. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *British Journal of Cancer*, 125, 04 2021.

[33] I. A. Mayer, V. G. Abramson, B. D. Lehmann, and J. A. Pietenpol. New strategies for triple-negative breast cancer—deciphering the heterogeneity. *Clinical cancer research*, 20(4):782–790, 2014.

[34] E. Montagna, P. Maisonneuve, N. Rotmensz, G. Cancello, M. Iorfida, A. Balduzzi, V. Galimberti, P. Veronesi, A. Luini, G. Pruneri, et al. Heterogeneity of triple-negative breast cancer: histologic subtyping to inform the outcome. *Clinical breast cancer*, 13(1):31–39, 2013.

[35] J. Morris. *Positive results : making the best decisions when you're at high risk for breast or ovarian cancer*. springer, 2011.

[36] P. Mullan and R. Millikan. Molecular subtyping of breast cancer: opportunities for new therapeutic approaches. *Cellular and molecular life sciences: CMLS*, 64(24):3219–3232, 2007.

[37] M. Neumann, D. King, I. Beltagy, and W. Ammar. Scispacy: Fast and robust models for biomedical natural language processing. *CoRR*, abs/1902.07669, 2019.

[38] J. O. Okoye. Molecular subtyping of triple negative breast cancer (tnbc): An approach to improving treatment response and survival outcome. *Medical Journal of Zambia*, 46(3):262–263, 2019.

[39] K. Polyak. Heterogeneity in breast cancer. *The Journal of clinical investigation*, 121:3786–8, 10 2011.

[40] S. Rontogianni, E. Synadaki, B. Li, M. C. Liefaard, E. H. Lips, J. Wesseling, W. Wu, and M. Altelaar. Proteomic profiling of extracellular vesicles allows for human breast cancer subtyping. *Communications biology*, 2(1):1–13, 2019.

[41] A. Rossanez, J. dos Reis, R. Torres, and H. De Ribaupierre. Kgen: A knowledge graph generator from biomedical scientific literature. *BMC Medical Informatics and Decision Making*, 20, 12 2020.

[42] T. Shibahara, C. Wada, Y. Yamashita, K. Fujita, M. Sato, A. Okamoto, and Y. Ono. Deep learning generates custom-made logistic regression models for explaining how breast cancer subtypes are classified. 2021.

[43] K. C. Souza, A. F. Evangelista, L. F. Leal, C. P. Souza, R. A. Vieira, R. L. Causin, A. Neuber, D. P. Pessoa, G. A. Passos, R. Reis, et al. Identification of cell-free circulating micrornas for the detection of early breast cancer and molecular subtyping. *Journal of Oncology*, 2019, 2019.

[44] R. G. Stein, D. Wollschläger, R. Kreienberg, W. Janni, M. Wischnewsky, J. Diessner, T. Stüber, C. Bartmann, M. Krockenberger, J. Wischhusen, et al. The impact of breast cancer biological subtyping on tumor size assessment by ultrasound and mammography-a retrospective multicenter cohort study of 6543 primary breast cancer patients. *BMC cancer*, 16(1):1–8, 2016.

[45] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, and J. Kang. Bern2: an advanced neural biomedical named entity recognition and normalization tool.

[46] K. Voduc, M. Cheang, S. Tyldesley, K. Gelmon, T. Nielsen, and H. Kennecke. Breast cancer subtypes and the risk of local and regional relapse. *Journal*

of clinical oncology : official journal of the American Society of Clinical Oncology, 28:1684–91, 03 2010.

[47] K.-J. Wang, B. Makond, K.-H. Chen, and K.-M. Wang. A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients. *Applied Soft Computing*, 20:15–24, 2014. Hybrid intelligent methods for health technologies.

[48] P. Welcsh. *The role of genetics in breast and reproductive cancers*. Springer, 2011.

[49] P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. Schütz, et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4):1–11, 2008.

[50] O. Yersal and S. Barutca. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology*, 5:412–424, 08 2014.

[51] L. Yin, J.-J. Duan, X.-W. Bian, and S.-c. Yu. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research*, 22(1):1–13, 2020.

[52] S. Yokoyama, T. Hamada, M. Higashi, K. Matsuo, K. Maemura, H. Kurahara, M. Horinouchi, T. Hiraki, T. Sugimoto, T. Akahane, S. Yonezawa, M. Kornmann, S. K. Batra, M. A. Hollingsworth, and A. Tanimoto. Predicted prognosis of patients with pancreatic cancer by machine learning. *Clinical Cancer Research*, 26(10):2411–2421, 2020.

APPENDICES

Appendix A

DATA SOURCES

A.1 Comprehensive molecular portraits of human breast tumours

We analysed primary breast cancers by genomic DNA copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays. Our ability to integrate information across platforms provided key insights into previously defined gene expression subtypes and demonstrated the existence of four main breast cancer classes when combining data from five platforms, each of which shows significant molecular heterogeneity. Somatic mutations in only three genes (TP53, PIK3CA and GATA3) occurred at ¿ 10% incidence across all breast cancers; however, there were numerous subtype-associated and novel gene mutations including the enrichment of specific mutations in GATA3, PIK3CA and MAP3K1 with the luminal A subtype. We identified two novel protein-expression-defined subgroups, possibly produced by stromal/microenvironmental elements, and integrated analyses identified specific signalling pathways dominant in each molecular subtype including a HER2/phosphorylated HER2/EGFR/phosphorylated EGFR signature within the HER2-enriched expression subtype. Comparison of basal-like breast tumours with high-grade serous ovarian tumours showed many molecular commonalities, indicating a related aetiology and similar therapeutic opportunities. The biological finding of the four main breast cancer subtypes caused by different subsets of genetic and epigenetic abnormalities raises the hypothesis that much of the clinically observable plasticity and heterogeneity occurs within, and not across, these major biological subtypes of breast cancer[26].

A.2 Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies

We pooled data from more than 10,000 cases of invasive breast cancer from 12 studies that had collected information on hormone receptor status, human epidermal growth factor receptor-2 (HER2) status, and at least one basal marker (cytokeratin [CK]5/6 or epidermal growth factor receptor [EGFR]) together with survival time data. Tumours were classified as luminal and nonluminal tumours according to hormone receptor expression. These two groups were further subdivided according to expression of HER2, and finally, the luminal and nonluminal HER2-negative tumours were categorised according to expression of basal markers. Changes in mortality rates over time differed by subtype. In women with luminal HER2-negative subtypes, mortality rates were constant over time, whereas mortality rates associated with the luminal HER2-positive and nonluminal subtypes tended to peak within 5 y of diagnosis and then decline over time. In the first 5 y after diagnosis the nonluminal tumours were associated with a poorer prognosis, but over longer follow-up times the prognosis was poorer in the luminal subtypes, with the worst prognosis at 15 y being in the luminal HER2-positive tumours. Basal marker expression distinguished the HER2-negative luminal and nonluminal tumours into different subtypes. These patterns were independent of any systemic adjuvant therapy[9].

A.3 Breast Cancer Consensus Subtypes: A system for subtyping breast cancer tumors based on gene expression

Breast cancer is heterogeneous in prognoses and drug responses. To organize breast cancers by gene expression independent of statistical methodology, we identified the Breast Cancer Consensus Subtypes (BCCS) as the consensus groupings of six different subtyping methods. Our classification software identified seven BCCS subtypes in a study cohort of publicly available data (n = 5950) including METABRIC, TCGA-BRCA, and data assayed by Affymetrix arrays. All samples were fresh-frozen from primary tumors. The estrogen receptor-positive (ER+) BCCS subtypes were: PCS1 (18%) good prognosis, stromal infiltration; PCS2 (15%) poor prognosis, highly proliferative; PCS3 (13%) poor prognosis, highly proliferative, activated IFN-gamma signaling, cytotoxic lymphocyte infiltration, high tumor mutation burden; PCS4 (18%) good prognosis, hormone response genes highly expressed. The ER- BCCS subtypes were: NCS1 (11%) basal; NCS2 (10%) elevated androgen response; NCS3 (5%) cytotoxic lymphocyte infiltration; unclassified tumors (9%). HER2+ tumors were heterogeneous with respect to BCCS[22].

A.4 Deep learning generates custom-made logistic regression models for explaining how breast cancer subtypes are classified

Breast cancer is the most frequently found cancer in women and the one most often subjected to genetic analysis. Nonetheless, it has been causing the largest number of women's cancer-related deaths. PAM50, the intrinsic subtype assay for breast cancer, is beneficial for diagnosis but does not explain each subtype's mechanism. Deep learning can predict the subtypes from genetic information more accurately than conventional statistical methods. However, the previous studies did not directly use deep learning to examine which genes associate with the subtypes. To reveal the mechanisms embedded in the PAM50 subtypes, we developed an explainable deep learning model called a point-wise linear model, which uses meta-learning to generate a custom-made logistic regression for each sample. We developed an explainable deep learning model called a point-wise linear model, which uses meta-learning to generate a custom-made logistic regression for each sample. Logistic regression is familiar to physicians, and we can use it to analyze which genes are important for prediction. The custom-made logistic regression models generated by the point-wise linear model used the specific genes selected in other subtypes compared to the conventional logistic regression model: the overlap ratio is less than twenty percent. Analyzing the point-wise linear model's inner state, we found that the point-wise linear model used genes relevant to the cell cycle-related pathways[42].

A.5 The androgen receptor as a surrogate marker for molecular apocrine breast cancer subtyping

The Androgen Receptor (AR) is a potential prognostic marker and therapeutic target in breast cancer. We evaluated AR protein expression in high-risk breast cancer treated in the adjuvant setting. Tumors were subtyped into luminal (ER+/Pg+-/AR+-), molecular apocrine (MAC, [ER-/PgR-/AR+]) and hormone receptor negative carcinomas (HR-negative, [ER-/PgR-/AR-]). Subtyping was evaluated with respect to prognosis and to taxane therapy. High histologic grade (p ¡ 0.001) and increased proliferation (p = 0.001) more often appeared in MAC and HR-negative than in luminal tumors. Patients with MAC had outcome comparable to the luminal group, while patients with HR-negative disease had increased risk for relapse and death. MAC outcome was favorable upon taxane-containing treatment; this remained significant upon multivariate analysis for overall survival (HR 0.31, 95%CI

0.13-0.74, interaction p = 0.035) and as a trend for time to relapse (p = 0.15). In conclusion, AR-related subtyping of breast cancer may be prognostic and serve for selecting optimal treatment combinations[27].

A.6 Proteomic profiling of extracellular vesicles allows for human breast cancer subtyping

Extracellular vesicles (EVs) are a potential source of disease-associated biomarkers for diagnosis. In breast cancer, comprehensive analyses of EVs could yield robust and reliable subtype-specific biomarkers that are still critically needed to improve diagnostic routines and clinical outcome. Here, we show that proteome profiles of EVs secreted by different breast cancer cell lines are highly indicative of their respective molecular subtypes, even more so than the proteome changes within the cancer cells. Moreover, we detected molecular evidence for subtype-specific biological processes and molecular pathways, hyperphosphorylated receptors and kinases in connection with the disease, and compiled a set of protein signatures that closely reflect the associated clinical pathophysiology. These unique features revealed in our work, replicated in clinical material, collectively demonstrate the potential of secreted EVs to differentiate between breast cancer subtypes and show the prospect of their use as non-invasive liquid biopsies for diagnosis and management of breast cancer patients[40].

A.7 PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers

Many methodologies have been used in research to identify the intrinsic subtypes of breast cancer commonly known as Luminal A, Luminal B, HER2-Enriched (HER2-E) and Basal-like. The PAM50 gene set is often used for gene expression-based subtyping; however, surrogate subtyping using panels of immunohistochemical (IHC) markers are still widely used clinically. Discrepancies between these methods may lead to different treatment decisions. We used the PAM50 RT-qPCR assay to expression profile 814 tumors from the GEICAM/9906 phase III clinical trial that enrolled women with locally advanced primary invasive breast cancer. All samples were scored at a single site by IHC for estrogen receptor (ER), progesterone receptor (PR), and Her2/neu (HER2) protein expression. Equivocal HER2 cases were confirmed by chromogenic in situ hybridization (CISH). Single gene scores by IHC/CISH were compared with RT-qPCR continuous gene expression values and intrinsic subtype assignment by the PAM50. High, medium, and low expression for ESR1, PGR, ERBB2, and proliferation were selected using quartile cut-points from the continuous RT-qPCR data across the PAM50 subtype assignments. ESR1, PGR, and ERBB2 gene expression had high agreement with established binary IHC cut-points (area under the curve (AUC) ¿= 0.9). Estrogen receptor positivity by IHC was strongly associated with Luminal (A and B) subtypes (92%), but only 75% of ER negative tumors were classified into the HER2-E and Basal-like subtypes. Luminal A tumors more frequently expressed PR than Luminal B (94% vs 74%) and Luminal A tumors were less likely to have high proliferation (11% vs 77%). Seventy-seven percent (30/39) of ER-/HER2+ tumors by IHC were classified as the HER2-E subtype. Triple negative tumors were mainly comprised of Basal-like (57%) and HER2-E (30%) subtypes. Single gene scoring for ESR1, PGR, and ERBB2 was more prognostic than the corresponding IHC markers as shown in a multivariate analysis. The standard immunohistochemical panel for breast cancer (ER, PR, and HER2) does not adequately identify the PAM50 gene expression subtypes. Although there is high agreement between biomarker scoring by protein immunohistochemistry and gene expression, the gene expression determinations for ESR1 and ERBB2 status was more prognostic[8].

A.8 Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures

Introduction Breast cancer subtyping and prognosis have been studied extensively by gene expression profiling, resulting in disparate signatures with little overlap in their constituent genes. Although a previous study demonstrated a prognostic concordance among gene expression signatures, it was limited to only one dataset and did not fully elucidate how the different genes were related to one another nor did it examine the contribution of well-known biological processes of breast cancer tumorigenesis to their prognostic performance.To address the above issues and to further validate these initial findings, we performed the largest meta-analysis of publicly available breast cancer gene expression and clinical data, which are comprised of 2,833 breast tumors. Gene coexpression modules of three key biological processes in breast cancer (namely, proliferation, estrogen receptor [ER], and HER2 signaling) were used to dissect the role of constituent genes of nine prognostic signatures.Using a meta-analytical approach, we consolidated the signatures associated with ER signaling, ERBB2 amplification, and proliferation. Previously published expression-based nomenclature of breast cancer 'intrinsic' subtypes can be mapped to the three modules, namely, the ER.sup.-.sup./HER2.sup.- .sup.(basal-like), the HER2.sup.+ .sup.(HER2-like), and the low- and high-proliferation ER.sup.+.sup./HER2.sup.- .sup.subtypes (luminal A and B). We showed that all nine prognostic signatures exhibited a similar prognostic performance in the entire dataset. Their prognostic abilities are due mostly to the detection of proliferation activity. Although ER.sup.- .sup.status (basal-like) and ERBB2.sup.+ .sup.expression status correspond to bad outcome, they seem to act through elevated expression of proliferation genes and thus contain only indirect information about prognosis. Clinical variables measuring the extent of tumor progression, such as tumor size and nodal status, still add independent prognostic information to proliferation genes. This meta-analysis unifies various results of previous gene expression studies in breast cancer. It reveals connections between traditional prognostic factors, expression-based subtyping, and prognostic signatures, highlighting the important role of proliferation in breast cancer prognosis[49].

A.9 Performance of three-biomarker immunohistochemistry for intrinsic breast cancer subtyping in the AMBER consortium

Background: Classification of breast cancer into intrinsic subtypes has clinical and epidemiologic importance. To examine accuracy of IHC-based methods for identifying intrinsic subtypes, a three-biomarker IHC panel was compared with the clinical record and RNA-based intrinsic (PAM50) subtypes. Automated scoring of estrogen receptor (ER), progesterone receptor (PR), and HER2 was performed on IHC-stained tissue microarrays comprising 1,920 cases from the African American Breast Cancer Epidemiology and Risk (AMBER) consortium. Multiple cores (1-6/case) were collapsed to classify cases, and automated scoring was compared with the clinical record and to RNA-based subtyping. Automated analysis of the three-biomarker IHC panel produced high agreement with the clinical record (93% for ER and HER2, and 88% for PR). Cases with low tumor cellularity and smaller core size had reduced agreement with the clinical record. IHC-based definitions had high agreement with the clinical record regardless of hormone receptor positivity threshold (1% vs. 10%), but a 10% threshold produced highest agreement with RNA-based intrinsic subtypes. Using a 10% threshold, IHC-based definitions identified the basal-like intrinsic subtype with high sensitivity (86%), although sensitivity was lower for luminal A, luminal B, and HER2-enriched subtypes (76%, 40%, and 37%, respectively). Three-biomarker IHC-based subtyping has reasonable accuracy for distinguishing basal-like from nonbasal-like, although additional biomarkers are required for accurate classification of luminal

A, luminal B, and HER2-enriched cancers. Epidemiologic studies relying on three-biomarker IHC status for subtype classification should use caution when distinguishing luminal A from luminal B and when interpreting findings for HER2-enriched cancers[4].

A.10 Concordance of clinical and molecular breast cancer subtyping in the context of preoperative chemotherapy response

ER, PR and HER2 status in breast cancer are important markers for the selection of drug therapy. By immunohistochemistry (IHC), three major breast cancer subtypes can be distinguished: Triple negative (TNIHC), [HER2+.sub.IHC] and [Luminal.sub.IHC] ([ER+.sub.IHC]/[HER2-.sub.IHC]). By using the intrinsic gene set defined by Hu et al. five molecular subtypes ([Basal.sub.mRNA], [HER2+.sub.mRNA], Luminal [A.sub.mRNA], Luminal [B.sub.mRNA] and Normal-[like.sub.mRNA]) can be defined. We studied the concordance between analogous subtypes and their prediction of response to neoadjuvant chemotherapy. We classified 195 breast tumors by both IHC and mRNA expression analysis of patients who received neoadjuvant treatment at the Netherlands Cancer institute for Stage II–III breast cancer between 2000 and 2007. The pathological complete remission (pCR) rate was used to assess chemotherapy response. The IHC and molecular subtypes showed high concordance with the exception of the [HER2+.sub.IHC] group. 60% of the [HER2+.sub.IHC] tumors were not classified as [HER2+.sub.mRNA]. The [HER2+.sub.IHC]/Luminal A or BmRNA group had a low response rate to a trastuzumab-chemotherapy combination with a pCR rate of 8%, while the [HER2+.sub.mRNA] group had a pCR rate of 54%. The Luminal [A.sub.mRNA] and Luminal [B.sub.mRNA] groups showed similar degrees of response to chemotherapy. Neither the PR status nor the endocrine responsiveness index subdivided the [ER+.sub.IHC] tumors accurately into Luminal [A.sub.mRNA] and Luminal [B.sub.mRNA] groups. Molecular subtyping suggests the existence of a [HER2+.sub.IHC]/[Luminal.sub.mRNA] group that responds poorly to trastuzumab-based chemotherapy. For [Luminal.sub.IHC] and triple [negative.sub.IHC] tumors, further subdivision into molecular subgroups does not offer a clear advantage in treatment selection[12].

A.11 New strategies for triple-negative breast cancer—deciphering the heterogeneity

Triple-negative breast cancer (TNBC) is a heterogeneous disease; gene expression analyses recently identified 6 distinct TNBC subtypes, each of which displays a unique biology. Exploring novel approaches for the treatment of these subtypes is critical, especially because the median survival for women with metastatic TNBC is less than 12 months, and virtually all women with metastatic TNBC ultimately will die of their disease despite systemic therapy. To date, not a single targeted therapy has been approved for the treatment of TNBC, and cytotoxic chemotherapy remains the standard treatment. In this review, the authors discuss recent developments in subtyping TNBC and the current and upcoming therapeutic strategies being explored in an attempt to target TNBC[33].

A.12 Breast cancer subtyping by immunohistochemistry and histological grade outperforms breast cancer intrinsic subtypes in predicting neoadjuvant chemotherapy response

Intrinsic subtypes are widely accepted for the classification of breast cancer. Lacking gene expression data, surrogate classifications based on immunohistochemistry (IHC) have been proposed. A recent St. Gallen consensus meeting recommends to use this surrogate intrinsic subtypes for predicting adjuvant chemotherapy resistance, implying that Surrogate Luminal A breast cancers should only receive endocrine therapy. In this study we assessed both gene expression based intrinsic subtypes as well as surrogate intrinsic subtypes regarding their power to predict neoadjuvant chemotherapy benefit. Single institution data of 560 breast cancer patients were reviewed. Gene expression data was available for 247 patients. Subtypes were determined on the basis of IHC, Ki67, histological grade, endocrine responsiveness, and gene expression, and were correlated with chemotherapy response and recurrence-free survival. In ER+/HER2- tumors, a high histological grade was the best predictor for chemotherapy benefit, both in terms of pCR (p = 0.004) and recurrence-free survival (p = 0.002). The gene expression based and surrogate intrinsic subtype based on Ki67 had no predictive or prognostic value in ER+/HER2- tumors. Histological grade, ER, PR, and HER2 were the best predictive factors for chemotherapy response in breast cancer. We propose to continue the conventional use of these markers[29].

A.13 The detection of ESR1/PGR/ERBB2 mRNA levels by RT-QPCR: a better approach for subtyping breast cancer and predicting prognosis

The molecular classification of breast cancer mainly focuses on ER, PR, and HER2 status detected by immunohistochemistry (IHC) analysis. To explore the clinical value of breast cancer classification based on gene-based diagnosis of the triple markers, we measured ESR1, PGR, and ERBB2 mRNA levels in 294 breast cancer patients by reverse transcription quantitative polymerase chain reaction (RT-QPCR), and examined their correlation with ER, PR, and HER2 status detected by IHC. We observed a significant positive correlation between the mRNA levels of the triple markers and their protein status (ESR1 vs. ER, Spearman's [rho] = 0.527, P = 2.3 x [10.sup.-22]; PGR vs. PR, Spearman's [rho] = 0.631, P = 5.1 x [10.sup.-34]; ERBB2 vs. HER2, Spearman's [rho] = 0.439, P = 3.0 x [10.sup.-15]). Furthermore, the subtypes determined by mRNA levels of the triple markers were significantly correlated to the subtypes determined based on their protein status (Spearman's [rho] = 0.342, P = 2.0 x [10.sup.-8]). Kaplan-Meier analysis showed that the subtypes determined by mRNA levels of the triple-marker could predict the disease-free survival (DFS) in breast cancer patients. Multivariate analysis showed that the predictive value of DFS could be confirmed for the subtypes determined by mRNA levels of the triple markers (HR = 2.285, P = 0.008) but not for those determined by their protein status. Taken together, our results suggest that the detection of ESR1/PGR/ERBB2 mRNA levels by RT-QPCR is a better approach for subtyping breast cancer and predicting the prognosis[13].

A.14 Triple-negative breast cancer molecular subtyping and treatment progress

Triple-negative breast cancer (TNBC), a specific subtype of breast cancer that does not express estrogen receptor (ER), progesterone receptor (PR), or human epidermal growth factor receptor 2 (HER-2), has clinical features that include high invasiveness, high metastatic potential, proneness to relapse, and poor prognosis. Because TNBC tumors lack ER, PR, and HER2 expression, they are not sensitive to endocrine therapy or HER2 treatment, and standardized TNBC treatment regimens are still lacking. Therefore, development of new TNBC treatment strategies has become an urgent clinical need. By summarizing existing treatment regimens, therapeutic drugs, and their efficacy for different TNBC subtypes and reviewing some new preclinical studies and targeted treatment regimens for TNBC, this paper aims to provide new ideas for TNBC treatment[51].

A.15 Comparison of breast cancer surrogate subtyping using a closed-system RT-qPCR breast cancer assay and immunohistochemistry on 100 core needle biopsies with matching surgical specimens

Routine clinical management of breast cancer (BC) currently depends on surrogate subtypes according to estrogen- (ER) and progesterone (PR) receptor, Ki-67, and HER2-status. However, there has been growing demand for reduced immunohistochemistry (IHC) turnaround times. The Xpert[R] Breast Cancer STRAT4* Assay (STRAT4)*, a standardized test for ESR1/PGR/MKi67/ERBB2 mRNA biomarker assessment, takes less than 2 hours. Here, we compared the concordance between the STRAT4 and IHC/SISH, thereby evaluating the effect of method choice on surrogate subtype assessment and adjuvant treatment decisions.In total, 100 formalin-fixed paraffin-embedded core needle biopsy (CNB) samples and matching surgical specimens for 98 patients with primary invasive BC were evaluated using the STRAT4 assay. The concordance between STRAT4 and IHC was calculated for individual markers for the CNB and surgical specimens. In addition, we investigated whether changes in surrogate BC subtyping based on the STRAT4 results would change adjuvant treatment recommendations. The overall percent agreement (OPA) between STRAT4 and IHC/SISH ranged between 76 and 99% for the different biomarkers. Concordance for all four biomarkers in the surgical specimens and CNBs was only 66 and 57%, respectively. In total, 74% of surgical specimens were concordant for subtype, regardless of the method used. IHC- and STRAT4-based subtyping for the surgical specimen were shown to be discordant for 25/98 patients and 18/25 patients would theoretically have been recommended a different adjuvant treatment, primarily receiving more chemotherapy and trastuzumab. A comparison of data from IHC/in situ hybridization and STRAT4 demonstrated that subsequent changes in surrogate subtyping for the surgical specimen may theoretically result in more adjuvant treatment given, primarily with chemotherapy and trastuzumab[23].

A.16 Molecular subtyping of breast cancer: opportunities for new therapeutic approaches

Evidence is accumulating that breast cancer is not one disease but many separate diseases. DNA microarray-based gene expression profiling has demonstrated subtypes with distinct phenotypic features and clinical responses. Prominent among the new subtypes is basal-like breast cancer, one of the intrinsic subtypes defined by negativity for the estrogen, progesterone, and HER2/neu receptors and positivity for cytokeratins-5/6. Focusing on basal-like breast cancer, we discuss how molecular technologies provide new chemotherapy targets, optimising treatment whilst sparing patients from unnecessary toxicity. Clinical trials are needed that incorporate long-term follow-up of patients with well-characterised tumour markers. Whilst the absence of an obvious dominant oncogene driving basal-like breast cancer and the lack of specific therapeutic agents are serious stumbling blocks, this review will highlight several promising therapeutic candidates currently under evaluation. Thus, new molecular technologies should provide a fundamental foundation for better understanding breast and other cancers which may be exploited to save lives[36]. (Part of a Multi-author Review)

A.17 Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data

Breast cancer is a highly heterogeneous disease. Subtyping the disease and identifying the genomic features driving these subtypes are critical for precision oncology for breast cancer. This study focuses on developing a new computational approach for breast cancer subtyping. We proposed to use Bayesian tensor factorization (BTF) to integrate multi-omics data of breast cancer, which include expression profiles of RNA-sequencing, copy number variation, and DNA methylation measured on 762 breast cancer patients from The Cancer Genome Atlas. We applied a consensus clustering approach to identify breast cancer subtypes using the factorized latent features by BTF. Subtype-specific survival patterns of the breast cancer patients were evaluated using Kaplan-Meier (KM) estimators. The proposed approach was compared with other state-of-the-art approaches for cancer subtyping. The BTF-subtyping analysis identified 17 optimized latent components, which were used to reveal six major breast cancer subtypes. Out of all different approaches, only the proposed approach showed distinct survival patterns (p ¡ 0.05). Statistical tests also showed that the identified clusters have statistically significant distributions. Our results showed that the proposed approach is a promising strategy to efficiently use publicly available multi-omics data to identify breast cancer subtypes[31].

A.18 Identification of cell-free circulating microRNAs for the detection of early breast cancer and molecular subtyping

Early detection is crucial for achieving a reduction in breast cancer mortality. Analysis of circulating cell-free microRNAs present in the serum of cancer patients has emerged as a promising new noninvasive biomarker for early detection of tumors and for predicting their molecular classifications. The rationale for this study was to identify subtype-specific molecular profiles of cell-free microRNAs for early detection of breast cancer in serum. Fifty-four early-stage breast cancers with 27 age-matched controls were selected for circulating microRNAs evaluation in the serum. The 54 cases were molecularly classified (luminal A, luminal B, luminal B Her2 positive, Her-2, triple negative). NanoString platform was used for digital detection and quantitation of 800 tagged microRNA probes and comparing the overall differences in serum microRNA expression from breast cancer cases with controls. We identified the 42 most significant (P [less than or equal to] 0.05,1.5-fold) differentially expressed circulating microRNAs in each molecular subtype for further study. Of these microRNAs, 19 were significantly differentially expressed in patients presenting with luminal A, eight in the luminal B, ten in luminal B HER 2 positive, and four in the HER2 enriched subtype. AUC is high with suitable sensitivity and specificity For the triple negative subtype miR-25-3p had the best accuracy Predictive analysis of the mRNA targets suggests they encode proteins involved in molecular pathways such as cell adhesion, migration, and proliferation. This study identified subtype-specific molecular profiles of cell-free microRNAs suitable for early detection of breast cancer selected by comparison to the microRNA profile in serum for female controls without apparent risk of breast cancer. This molecular profile should be validated using larger cohort studies to confirm the potential of these miRNA for future use as early detection biomarkers that could avoid unnecessary biopsy in patients with a suspicion of breast cancer[43].

A.19 Heterogeneity of triple-negative breast cancer: histologic subtyping to inform the outcome

This study assesses outcome in terms of disease-free survival (DFS) and overall survival (OS) of special types of triple-negative breast cancer (TNBC). We identified 8801 women with first primary nonmetastatic breast cancer operated on at the European Institute of Oncology between 1997 and 2005. Of these patients, 781 consecutive patients with immunohistochemically defined TNBC were selected for the analyses. We explored patterns of recurrence by histologic type. Median follow-up was 5.7 years (range 0-13 years). The 5-year DFS was 77% for TNBC, 68% for human epidermal growth factor receptor 2 (HER2)-positive breast cancer, and 84% and 95% for luminal B and luminal A breast cancer, respectively. From 781 TNBC subtypes, 693 cases (89%) were classified as ductal not otherwise specified (NOS) (invasive ductal carcinoma [IDC]), 29 were classified as apocrine (3.7%), 18 (2.3%) were classified as lobular, 10 (1.2%) were classified as adenoid cystic, and 10 (1.2%) were classified as metaplastic. Five-year DFS and OS were 77% and 84% for patients with ductal carcinoma, 56% and 89% for patients with metaplastic carcinoma, and both 5-year DFS and OS were 100% for patients with adenoid cystic and medullary carcinomas, respectively. Distinct prognostic implications may derive from the specific histotype of TNBC. The identification of these special types has a significant clinical utility and should be considered in therapeutic algorithms[34].

A.20 An aptamer-based probe for molecular subtyping of breast cancer

Molecular subtyping of breast cancer is of considerable interest owing to its potential for personalized therapy and prognosis. However, current methodologies cannot be used for precise subtyping, thereby posing a challenge in clinical practice. The aim of the present study is to develop a cell-specific single-stranded DNA (ssDNA) aptamer-based fluorescence probe for molecular subtyping of breast cancer. Cell-SELEX method was utilized to select DNA aptamers. Flow cytometry and confocal microscopy were used to study the specificity, binding affinity, temperature effect on the binding ability and target type analysis of the aptamers. In vitro and in vivo fluorescence imaging were used to distinguish the molecular subtypes of breast cancer cells, tissue sections and tumor-bearing mice. Six SK-BR-3 breast cancer cell-specific ssDNA aptamers were evolved after successive in vitro selection over 21 rounds by Cell-SELEX. The Kd values of the selected aptamers were all in the low-nanomolar range, among which aptamer sk6 showed the lowest Kd of 0.61 +/- 0.14 nM. Then, a truncated aptamer-based probe, sk6Ea, with only 53 nt and high specificity and binding affinity to the target cells was obtained. This aptamer-based probe was able to 1) differentiate SK-BR-3, MDA-MB-231, and MCF-7 breast cancer cells, as well as distinguish breast cancer cells from MCF-10A normal human mammary epithelial cells; 2) distinguish HER2-enriched breast cancer tissues from Luminal A, Luminal B, triple-negative breast cancer tissues, and adjacent normal breast tissues (ANBTs) in vitro; and 3) distinguish xenografts of SK-BR-3 tumor-bearing mice from those of MDA-MB-231 and MCF-7 tumor-bearing mice within 30 min in vivo. The results suggest that the aptamer-based probe is a powerful tool for fast and highly sensitive subtyping of breast cancer both in vitro and in vivo and is also very promising for the identification, diagnosis, and targeted therapy of breast cancer molecular subtypes[30].

A.21 St Gallen 2015 subtyping of luminal breast cancers: impact of different Ki67-based proliferation assessment methods

Ki67 has been proposed as prognostic proliferation marker in luminal breast cancer (BC), but little is known on the influence of Ki67 assessment methods on subtyping into luminal A- and B-like tumors. Our aim was to study the influence of different Ki67-labeling index (Ki67-LI) assessment methods on the proportion of BCs classified as luminal A-like. 280 early BCs were subtyped according to the St Gallen 2015 definitions into 71 % luminal (HER2 negative), 6 % luminal B-like (HER2 positive), 13 % triple negative, 1 % HER2 positive (nonluminal), and 9 % special type. Digitized whole slides were counted manually on the screen. We used nine defined counting methods to assess the Ki67-LI (including the International Ki67 in Breast Cancer Working Group recommendations), and compared the resulting medians and the proportions of cancers classified as luminal A-like according to the formerly used cut-off

¡20 %. Methods assessing hot spots and tumor periphery resulted in significantly higher Ki67-LI medians than those measuring an average proliferation (27.45 % vs 16.96 %, p ¡ 0.0001). Substantially lower median Ki67-LI were found when assessing 1020 compared to counting 100, 200, 300 cells (17.65 vs 33%, vs 28 %, vs 24.33 %, respectively; p ¡ 0.0001), or 510 cells (20.59 %, p = 0.019). Applying a standard Ki67-LI cut-off ¡20 % to define low proliferation for all methods, the proportion of luminal A-like cancers varied between 13 and 44 %. The proportion of BCs classified as luminal A-like is highly influenced by the Ki67-LI assessment method. As a consequence, the selection of a specific Ki67-LI assessment method may have a direct effect on the proportion of patients considered having low-risk disease and thus influence therapeutic decision making. This calls for a standardized assessment method[15].

A.22 Molecular Subtyping of Triple Negative Breast Cancer (TNBC): An approach to improving treatment response and survival outcome

Molecular subtyping of triple-negative breast cancers (TNBCs) via gene expression profiling is essential for understanding the molecular essence of this heterogeneous disease and for guiding individualized treatment. We aim to devise a clinically practical method based on immunohistochemistry (IHC) for the molecular subtyping of TNBCs. By analyzing the RNA sequencing data on TNBCs from Fudan University Shanghai Cancer Center (FUSCC) (n = 360) and The Cancer Genome Atlas data set (n = 158), we determined markers that can identify specific molecular subtypes. We performed immunohistochemical staining on tumor sections of 210 TNBCs from FUSCC, established an IHC-based classifier, and applied it to another two cohorts (n = 183 and 214). We selected androgen receptor (AR), CD8, FOXC1, and DCLK1 as immunohistochemical markers and classified TNBCs into five subtypes based on the staining results: (a) IHC-based luminal androgen receptor (IHC-LAR; AR-positive

[+]), (b) IHC-based immunomodulatory (IHC-IM; AR-negative [-], CD8+), (c) IHC-based basal-like immune-suppressed (IHC-BLIS; AR-, CD8-, FOXC1+), (d) IHC-based mesenchymal (IHC-MES; AR-, CD8-, FOXC1-, DCLK1+), and (e) IHC-based unclassifiable (AR-, CD8-, FOXC1-, DCLK1-). The k statistic indicated substantial agreement between the IHC-based classification and mRNA-based classification. Multivariate survival analysis suggested that our IHC-based classification was an independent prognostic factor for relapse-free survival. Transcriptomic data and pathological observations implied potential treatment strategies for different subtypes. The IHC-LAR subtype showed relative activation of HER2 pathway. The IHC-IM subtype tended to exhibit an immune-inflamed phenotype characterized by the infiltration of CD8+ T cells into tumor parenchyma. The IHC-BLIS subtype showed high expression of a VEGF signature. The IHC-MES subtype displayed activation of JAK/STAT3 signaling pathway. We developed an IHC-based approach to classify TNBCs into molecular subtypes. This IHC-based classification can provide additional information for prognostic evaluation. It allows for subgrouping of TNBC patients in clinical trials and evaluating the efficacy of targeted therapies within certain subtypes[38].

A.23 Development of an absolute assignment predictor for triple-negative breast cancer subtyping using machine learning approaches

Triple-negative breast cancer (TNBC) heterogeneity represents one of the main obstacles to precision medicine for this disease. Recent concordant transcriptomics studies have shown that TNBC could be divided into at least three subtypes with potential therapeutic implications. Although a few studies have been conducted to predict TNBC subtype using transcriptomics data, the subtyping was partially sensitive and limited by batch effect and dependence on a given dataset, which may penalize the switch to routine diagnostic testing. Therefore, we sought to build an absolute predictor (i.e., intra-patient diagnosis) based on machine learning algorithms with a limited number of probes. To that end, we started by introducing probe binary comparison for each patient (indicators). We based the predictive analysis on this transformed data. Probe selection was first involved combining both filter and wrapper methods for variable selection using cross-validation. We tested three prediction models (random forest, gradient boosting [GB], and extreme gradient boosting) using this optimal subset of indicators as inputs. Nested cross-validation consistently allowed us to choose the best model. The results showed that the fifty selected indicators highlighted the biological characteristics associated with each TNBC subtype. The GB based on this subset of indicators performs better than other models[7].

A.24 The impact of breast cancer biological subtyping on tumor size assessment by ultrasound and mammography-a retrospective multicenter cohort study of 6543 primary breast cancer patients

Mammography and ultrasound are the gold standard imaging techniques for preoperative assessment and for monitoring the efficacy of neoadjuvant chemotherapy in breast cancer. Maximum accuracy in predicting pathological tumor size non-invasively is critical for individualized therapy and surgical planning. We therefore aimed to assess the accuracy of tumor size measurement by ultrasound and mammography in a multicentered health services research study. We retrospectively analyzed data from 6543 patients with unifocal, unilateral primary breast cancer. The maximum tumor diameter was measured by ultrasound and/or mammographic imaging. All measurements were compared to final tumor diameter determined by postoperative histopathological examination. We compared the precision of each imaging method across different patient subgroups as well as the method-specific accuracy in each patient subgroup. Overall, the correlation with histology was 0.61 for mammography and 0.60 for ultrasound. Both correlations were higher in pT2 cancers than in pT1 and pT3. Ultrasound as well as mammography revealed a significantly higher correlation with histology in invasive ductal compared to lobular cancers (p ¡ 0.01). For invasive lobular cancers, the mammography showed better correlation with histology than ultrasound (p = 0.01), whereas there was no such advantage for invasive ductal cancers. Ultrasound was significantly superior for HR negative cancers (p ¡ 0.001). HER2/neu positive cancers were also more precisely assessed by ultrasound (p ¡ 0.001). The size of HER2/neu negative cancers could be more accurately predicted by mammography (p ¡ 0.001). This multicentered health services research approach demonstrates that predicting tumor size by mammography and ultrasound provides accurate results. Biological tumor features do, however, affect the diagnostic precision[44].