Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 13-005

Integration of Clinical and Genomic data: a Methodological Survey

Sanjoy Dey, Rohit Gupta, Michael Steinbach, and Vipin Kumar

February 20, 2013

Integration of Clinical and Genomic data: a Methodological Survey

Sanjoy Dey[1*], Rohit Gupta[1], Michael Steinbach[1], Vipin Kumar[1]

[1]Department of Computer Science, University of Minnesota, MN 55455, USA.

Abstract: Human diseases are inherently complex and governed by the complicated interplay of several underlying factors. Clinical research focuses on behavioral, demographic and pathology information, whereas molecular genomics focuses on finding underlying genetic and genomic factors in genomic data collected on mRNA expression, proteomics, biological networks, and other microbiological features. However, each of these clinical and genomic datasets contains information only about one particular aspect of a complex disease, rather than covering all of the several complicated underlying risk factors. This has led to a new area of research that integrates both clinical and genomic data and aims to extract more information about diseases by considering not only all the various factors, but also the interactions among those factors, which cannot be captured by clinical and genomic studies that are performed independently of each other. Although initial efforts have already been made to develop such integrative modeling of the clinical and genomic data to shed light on the biological mechanism of the diseases, the research field is still in a rudimentary stage. In this review article, we survey the general issues, challenges and current work of clinicogenomic studies. We also summarize the current state of the field and discuss some possibilities for future work.

1 Background and Motivation

Until the last decade, traditional clinical care and management of complex diseases mainly relied on different clinico-pathological data, such as signs and symptoms, demographic data, pathological lab test results, and medical images. In addition, efforts have been made to capture genetic factors by maintaining the family history of patients. The effect of such clinical and histo-pathological markers is assessed by cohort based studies conducted on large populations (Szklo 1998) and the knowledge obtained from these studies is summarized as clinical guidelines for the diagnosis, prognosis, monitoring and treatment of human disease, e.g., NPI (Galea, Blamey et al. 1992) and Adjuvant! Online (Goldhirsch, Coates et al. 2006) for breast cancer and palmOne (Blumberg 2004) for prostate cancer. However, this approach still falls short. For example, there are adverse drug reactions for some patients who have risk factors similar to those patients who have been cured by the same therapeutic treatment. This issue stems from the strategy of 'one drug fits all' and motivates the need to improve on conclusions drawn from cohort-based studies so that the underlying mechanism of complex diseases can be understood at the individual patient level.

The recent advancement of high-throughput technology has led to an abundance of information for each individual at the micro-molecular level. A myriad of genetic, genomic and metabolomics data have been collected to capture different aspects of cell mechanism that shed light on human physiology. Examples include SNPs, which provide information about the genetic polymorphism of an individual; gene expressions, which measure transcription; and protein and metabolite abundance, which captures protein abundance and post-translational modifications. These high-throughput datasets have helped answer some complex biological questions for different diseases, such as assessing the prognosis effect (Sotiriou and Piccart 2007),(Driouch, Landemaine et al. 2007),(Potti, Mukherjee et al. 2006),(Garber, Troyanskaya et al. 2001), epistasis effects on diseases (Anastassiou 2007), and discovering new sub-phenotypes of complex diseases (Golub, Slonim et al. 1999),(Alizadeh, Eisen et al. 2000),(Bhattacharjee, Richards et al. 2001). The use of genetic information in epidemiology helped design effective diagnostics, new therapeutics, and novel drugs which have led to the recent era of *personalized medicine* (genomic medicine) (Stephenson, Smith et al. 2005), (Edén, Ritz et al. 2004), (Teschendorff, Naderi et al. 2006). However, these genetic factors *alone* cannot explain all the intricacies of complex diseases. For example, the incidences of cancer vary widely among different countries due to the environmental factors, even for the same ethnic groups, when they migrate from one country to another (Redmond Jr 1970), (Weinberg 2007).

In recent studies (Schadt 2009, Eichler, Flint et al. 2010), it has been hypothesized that most complex diseases are caused by the *combined* effects of many diverse factors, including different genetic, genomic, behavioral factors and environmental effects. For example, cancer, which is the most widely studied disease phenotype in last few decades, is extremely heterogeneous. Different clinical endpoints of cancer, such as the idiosyncrasy of individual tumors, the survival rate of cancer patients after chemotherapy or surgical treatment, development of metastasis, and the effectiveness of drug therapy are governed by different risk factors including multiple mutations of genetic factors (e.g., RAS, RTK, TGF-β, Wnt/signaling pathways), behavioral factors (e.g. tobacco exposure, diet, lifestyle) (Weinberg 2007), long-time environmental effects (e.g., stresses, temperature, radiation, oxygen tensions, hydration and tonicity, micro- and macro-nutrients, toxins) (Loscalzo, Kohane et al. 2007) and inherent germline variations (e.g. BRCA1/2) (West, Ginsburg et al. 2006). Therefore, clinico-pathological and genomic datasets capture different effects of all such diverse factors on complex diseases in a complementary manner rather than a supplementary nature. Using the two diverse perspectives provided by both types of data can potentially reveal disease complexities in greater details.

In addition, the individual effects of each of the clinicogenomic factors on disease predisposition can be small and thus can remain undetected by most disease association techniques performed on individual datasets. However, interactions among those individual factors may be responsible for increasing the risk of complex disease (Anastassiou 2007),(Loscalzo, Kohane et al. 2007). For example, neither a gene nor an environmental factor like tobacco use may be significantly associated with lung cancer by itself, but together they can increase the risk significantly (Zhou, Liu et al. 2002). In a more complicated scenario, a complex genetic network can evolve dynamically under various environmental factors (Schadt 2009). This phenomenon is true even for Mendelian disease with monogenic disorder like sickle cell disease, where single different phenotypes were observed based on environmental effects (Kato, Gladwin et al. 2007). Besides interactions, there may be other types of relationships such as causal relationships between two types of markers (Lê Cao, Meugnier et al. 2010). For example, some pathological variables such as PSA, can also have upstream genetic influence (Singh, Febbo et al. 2002). In this case, the individual factors coming from different datasets may not be a strong biomarker; but rather the relationships inherent among them, including interaction, can act as potential biomarkers.

Leveraging such wider relationships including interactions, correlation and casualty among the genomic, pathological, environmental, and behavioral factors is important for understanding the nature of diseases. It will also assist in making better clinical decisions. For example, surgery can be avoided if some causative genomic markers of the tumor can be targeted in the early stage of breast cancer. Note that if the association of clinical and genomic factors with the disease phenotype is assessed independently, such deeper levels of relationships among data sources cannot be discovered. It is essential to build integrative models considering both genomic and clinical variables simultaneously being cognizant of the interaction, redundancy, and correlation among those clinical and genomic data (Schadt 2009). This has led to an emerging research area of integrative studies of clinical and genomic data, which we will refer as *clinico-genomic integration*. In this review, we survey not only different issues and challenges existing in such clinico-genomic integrative studies, but also different approaches that aimed to address those issues. Finally, we conclude with a general discussion on future research directions in this topic.

1.1 What is clinicogenomic integration?

Clinicogenomic integration means building models by integrating clinical and genomic data. Clinical data refers to a broad category of patient's pathological, behavioral, demographic, familial, environmental and medication history, while genomic data refers to any kind of patient's genetic information including SNPs, gene expression, protein and metabolite profiles [Figure 2]. More specifically, the clinicogenomic studies should have at least one clinical dataset and one genomic dataset for a group of people who are assessed for an outcome of a phenotype of a disease. Furthermore, we survey only integrative models with an emphasis on biomarker discovery. Therefore, each sample of datasets is assessed for a particular disease phenotype. The phenotype can be either binary class labels such as cancer vs. no cancer, tumor vs. normal tissue samples, metastasis vs. non-recurrent cancer, or continuous variables, e.g., the survival time after chemotherapy or other types of therapeutic treatments. Achieving the goal of biomarker discovery requires identifying the clinical and genomic features from the data that are significantly associated with the disease phenotype.

1.2 Different aspects of clinicogenomic studies

Integration of diverse biomedical datasets is a vast research topic and has been studied widely in many different domains. Although some initial efforts have been made by researchers for clinicogenomic integration, most of these studies are scattered throughout the literature and were developed from a clinical perspective for different disease phenotypes with their own limitations and advantages. Moreover, the issues and challenges related to this field are not yet well understood. In this article, we first identify the overall issues and challenges of this field with an emphasis on the methodological perspective (Section 2) and then discuss how the existing clinicogenomic methods address these challenges (Section 3). In particular, we categorize existing methods from many perspectives: stage of integration (Section 3.1), how disparate dimensionality is addressed (Section 3.2), and disease heterogeneity (Section 3.3). We also discuss the different goals that these studies try to achieve (Section 4) and the validation techniques used in each category (Section 5). Finally, we review several multi-site clinicogenomic models (Section 6) and then conclude with discussion (Section 7). Moreover, the scope of the article is the integrative model development of clinical and genomic data, rather than the simple incorporation of genomic data into clinical practice for designing genomic medicine.

There are some existing articles that focus on few aspects of the clinicogenomic studies (Table 1). Boulesteix et al. (Boulesteix and Sauerbrei 2011) performed a recent survey on how to validate the additional predictive power of genomic markers over traditional clinical variables with a focus on external data. Unlike this review, they did not aim at reviewing all types of predictive clinicogenomic models and the challenges these studies address. Correa et al. 2010 (Correa, Adali et al. 2010) partially reviewed the integration approaches that find relationships between datasets measured by correlation while Thomas et al. (Oelker and Boulesteix) mainly studied methods aiming to find interactions between genes and environmental factors. There are several other reviews on studies that integrate diverse genetic, genomic, proteomic, metabolomics, interactome, phylogeneic and phenome data (abbreviated as 'omic' data) (Hamid, Hu et al. 2009, Tsiliki and Kossida 2011, Bebek, Koyutürk et al. 2012). However, they do not cover the integration of clinical data with genomic datasets. Although they fall out of scope of our survey, they are included in Table 1 since they discuss some of the generic challenges common in any integrative study. To the best of our knowledge, currently there is no study which aims at reviewing integrative approaches combining genomic and clinical/environmental data from a methodological perspective.

2 Issues and challenges in integrating clinical and genomic data

The integration of genomic and clinical studies is difficult as the two fields have different perspectives. Several key technical challenges are described below.

1. **Difference in nature of the datasets**: As the datasets being integrated are collected from two different perspectives, the difference in the nature of the data being integrated creates several challenges for developing integrative models. First, clinical data are usually record-based where each patient can be represented by a record of clinical variables. On the other hand, genetic and genomic data sets vary widely in terms of formats. Besides record-based data, there are many network-based datasets where the relationships among several biomolecular entities are represented as features. Second, clinical variables are available in diverse data types such as text, categorical, and numeric values, but on the other hand genomic variables are mostly numeric. Third, some data sets may contain structure, e.g., measurements across time or across a genetic sequence, that are not present in others. Fourth, genomic and genetic datasets are high-dimensional in comparison with clinical data which often contains ~10-20 variables. Fifth, genetic and genomic data contains a higher level of missing values because of technological issues (Ioannidis 2005). In contrast, clinical data are easy and inexpensive to collect, and so contain fewer missing values. Integrating variables with such different formats, types, structure, dimensionality, and missing values is a challenging problem in the data mining and machine learning domain.
2. **Statistical significance:** The high-dimensionality of genomic datasets combined with low sample size poses challenge for finding statistically significant biomarkers (classical statistical n<<p problem) []. Combining such high dimensional features with low dimensional clinical data creates a further challenge for statistical and data mining methods. Even after pre-selecting significant genes, genomic features (which usually number a few hundred) still dominate the traditional clinical variables in number (which usually are around ~10-20) due to the over-fitting problem. Unless the whole experimental setup is performed cautiously, the clinical markers can be lost in the vast regime of genomic data and thus the

predictive power of the genomic data could be overestimated (Boulesteix, Porzelius et al. 2008),(Tibshirani, Efron et al. 2002).

3. **Different biases and assumptions:** Since the corresponding datasets are collected independently, the biases and assumptions of each of the data sets being integrated may be different due to the difference in experimental designs and protocols. For example, since clinical variables are gathered more systematically over a large period of time intervals, they contain less noise. In addition, they are validated rigorously by numerous epidemiology studies (Boulesteix, Porzelius et al. 2008, Truntzer, Maucort-Boulch et al. 2008, Obulkasim, Meijer et al. 2011). Also, clinical data are cheap and easy to collect (Obulkasim, Meijer et al. 2011). In contrast, gene expression data (and also other genomic data) are less reproducible over independent cohorts because of the high noise, different experimental biases, high-dimensionality and small sample sizes of the microarray datasets (Ein-Dor, Kela et al. 2005, Ein-Dor, Zuk et al. 2006, Naderi, Teschendorff et al. 2006, Chuang, Lee et al. 2007). Integrative studies need to be aware of such differing degrees of information present in different datasets. Otherwise, the role of clinical variables may be underestimated in the prognostic model when compared with noisy genomic variables (Truntzer, Maucort-Boulch et al. 2008).
4. **Heterogeneity:** Most complex diseases are heterogeneous in nature, i.e., patients with a particular disease may form different subgroups and factors appropriate for one subgroup may not apply to another. For example, different subsets of the population are known to have different biomarkers for the same disease (McClellan and King 2010), due to different pathways playing a role in the same disease, or due to the same pathway playing a different role in subjects from different ethnicities (Kelley and Ideker 2005). It is important to find such subgroups for integrative studies, because the effect of one type of variable in a subspace can be explained when integrated with other types of data, but the variable would be treated as a confounding factor otherwise.
5. **Interpretability:** Another generic challenge for biomarker studies is that the obtained model has to be interpretable, i.e., the effect of individual markers on the disease phenotype must be identified. Otherwise, the domain experts cannot use the potential risk factors for further validations. Similarly, clinicogenomic integrative models must be easily interpretable to be treated as biomarkers. For example, if the original genes or clinical variables cannot be interpreted by the model, then drugs cannot be designed targeting some genes or the proteins encoded by those genes.
6. **Finding relationships between datasets:** The relationship between clinical and genomic data may also be different. The predictive model achieves benefits over the individual studies if the two datasets are complementary in nature, but not supplementary in nature. Besides such broad relationships among overall datasets, relationships can be present among different types of individual markers. For example, correlation between genomic data and several clinical variables has been observed in many domains including cancer research (Singh, Febbo et al. 2002), (Sotiriou, Wirapati et al. 2006), (Kumar, Grigorakis et al. 2005) and neuroscience (Correa, Adali et al. 2010). There may exist other types of relationships including interactions (Oelker and Boulesteix) and causal (Boulesteix, Porzelius et al. 2008, Lê Cao, Meugnier et al. 2010) relationships among different types of variables as described earlier. Finding and leveraging all the relationships between clinical and genomic data poses a big challenge for integrative study.

Some of the technical challenges (the disparate natures of the data) described above are quite general and are applicable for any type of integrative studies in any domain, while some others (heterogeneity and interpretability) are applicable mainly for biomarker discovery. Besides these generic challenges, clinicogenomic integration also faces challenges (statistical significance, different amount of information) that are specific to the domain. It is difficult to address all these challenges. In fact, most of the clinicogenomic studies aim to address only a few challenges described earlier. Several of them were motivated by the integrative models from different research communities including biostatistics, data mining, and machine learning, which can handle the generic challenges mentioned above. Of course, many of them were further modified; sometimes completely new methods were designed to address the specific challenges of clinicogenomic data integration.

Besides these technical challenges, there are several domain challenges as well. First, there are differences in the terminologies used in epidemiology and genetics studies, even for the same topic. For example, association studies (genetics) and case-control studies (epidemiology) deal with the same concept of finding causative factors of diseases. (Maojo and Martin-Sanchez 2004). This makes the automatic extraction of information from health care and genetics data difficult. Second, the genomic data have been collected from a research perspective solely according to solid scientific theories and models. However, the health care data is collected slowly over a longer period of time in a retrospective manner, from different sources spanning broad areas such as medical observations, patient management data, healthcare providers, doctor's note, and patient's life history. Thus, clinical data collected from electronic medical records [EMR] may contain redundant information, which has to go through several preprocessing steps to extract useful information about a patient that can be integrated later with genomic data. Lastly, the privacy issue related to the healthcare domain (Hristidis 2009), (Meingast, Roosta et al. 2008) creates a serious bottleneck to the availably of clinical data. The data collection related challenges require several preprocessing steps such as building data warehouse integrating multiple sources of data, extracting information from them, text mining and natural language processing(NLP). Such data collection and preprocessing steps are out of the scope of this review, since the focus of this review is on developing integrative models.

3 Different types of integration

In this section, we identify and discuss several aspects of integrative model development. Furthermore, we will categorize the existing clinicogenomic studies based on those aspects and discuss how they address different challenges as described in the last section. Again, some of these categorizations and the corresponding studies which address the general challenges are applicable for any domain; while some address the specific challenge of clinicogenomic integration. The main goal of the review is to analyze the clinicogenomic models from a methodological perspective. Table 2 contains the overall summary of all clinicogenomic models based on these taxonomies.

3.1 Stages of data integration:

Integration of multiple heterogeneous datasets in general can be performed in several stages. For example, either individual datasets can be integrated first before developing any model or decisions coming from models built on each dataset independently can be integrated. Alternatively, each dataset can be transformed to a common intermediate structure such as graph or kernel and then these structures can be merged before developing models. Pavlidis et al. (Pavlidis, Weston et al. 2001) performed seminal work on these three types of integration and called them early, late and intermediate integration respectively. We will categorize all clinicogenomic integration into these three broad conceptual categories.

3.1.1 Early integration:

In general, early integrative approaches merge the independent data sources together before performing any kind of data analysis. In a simplistic case, the individual data matrices are simply augmented into a larger matrix if both of the datasets have same set (or subset) of samples. Thus, the integration of the individual datasets, which are clinical and genomic data in our case, is performed in an early stage of the overall analysis. Once the combined data matrix is prepared, any types of models can be developed based on the three goals of the clinicogenomic studies described in section 4.

The unique assumption of this type of integration is that both of the datasets are similar in nature, i.e., most of the properties of the datasets such as data type, formats, structure, dimensionality are either similar or preprocessed to be as similar as possible. Otherwise, a significant amount of preprocessing such as dimensionality reduction, missing value imputation and data discretization is required before integrating individual datasets.

Advantage: Early integration is the simplest approach, since any standard model can be applied on the integrated dataset to achieve any of the objectives. Therefore, most of the clinicogenomic studies fall in this category []. Moreover, they can preserve any kind of inter-data relationships. For example, if some clinical and genomic variables are correlated, the model developed after data integration can take the correlation structure into account.

Disadvantage: Early integration loses the individual properties of each dataset such as the structure and the different degree of information when merged together into an augmented dataset. The dimensionality of the augmented dataset also increases, thus the model may also suffer from high dimensionality and low statistical significance of the obtained result.

3.1.2 Late integration:

Late integration first develops predictive models separately for each of the individual data sources and then merges the individual decisions of all predictive models into a final score as the prediction of outcome variable. As opposed to early integration, this type of integration actually merges the classifier decision rather than original dataset. The main assumption of late integration is that the individual datasets are independent and there is no inter-datasets relationship.

The biggest challenge of late integration is how to merge the decision of classifiers obtained from individual datasets. Several strategies like majority voting, linear aggregation and weighted average have been applied for this purpose. For example, two breast cancer studies conducted by Campone et al. 2008 (Campone, Campion et al. 2008) and Silhava et al. 2009 (Šilhavá and Smrz 2009) simply summed up the individual decision coming from genomic and clinical data. Campone et al. applied the Cox regression model to summarize the topmost 15 discriminating genes into a single genomic score and then added it to the traditional clinical score of breast cancer, NPI to get the final score for assessing the effect of adjuvant chemotherapy. On the other hand, Silhava et al.(Šilhavá and Smrz 2009) applied two different predictive models: logistic regression, and BionomialBoosting(BB)(Buhlmann and Hothorn 2007) to get the genomic and clinical score respectively before summing up.

However, simple summation is not always appropriate because the contribution of the individual data sources to the overall clinicogenomic model may be different. Alternatively, the contribution from each individual dataset towards the disease phenotype can be assessed and the scores obtained from the individual models can be weighted accordingly. For example, Futschik et al, 2003 (Futschik, Sullivan et al. 2003) used parameterized learning for merging the individual decisions into the final decision. In order to integrate the individual decisions of the clinical (by Bayesian classifier) and genomic data (by evolution fuzzy artificial neural network (EFuNN (Kasabov 2001)), a modular hierarchical model was introduced based on two levels of parameters for assessing the confidence of the decisions of the two predictive models towards the class label and adjusting the class bias. Furthermore, they also tested statistical independency of the outputs of two independent models using the mutual information (Cover and Thomas 2006), which is the important assumption for late integration. In a more complicated scenario, with many datasets being integrated, the more general problem arises when some of the models built on individual datasets produce binary class decisions and some of the predictive models generate continuous-valued scores. Several approaches including majority vote, and its more generic version called consensus learning (Gao, Fan et al. 2009) have been studied in many other domains such as image processing, social networks domain.

Advantage: The individual structure and the nature of each dataset are preserved in late integration, since model is developed on each dataset separately. Moreover, different models can be used for different datasets depending on the individual nature of the datasets. Late integration is particularly useful when each of the datasets is completely heterogeneous, i.e., the datasets cannot be transformed into a common format for integration.

Disadvantage: Late integration misses any kind of possible relationship like correlation, interactions present among the datasets. Moreover, late integration generates a different hypothesis for each of the datasets as opposed to a single hypothesis for the integrated dataset. Interpretation and validation of these different types of hypothesis is not trivial.

3.1.3 Intermediate integration:

Early and late integration are opposite in nature in terms of their advantages and disadvantages. Intermediate integration tries to overcome the limitations of both approaches. It first represents each dataset with a common structure, such as a graph or kernel, and then merges these representations before developing any models.

Therefore, it generates one hypothesis, but can retain the structure of each data set and take in account the possible relationships between the datasets to some extent. The main assumption of this approach is that there is an appropriate intermediate representation for each dataset preserving the individual properties of that dataset and the intermediate representations can be combined easily.

Kernel based intermediate integration has become the most popular technique for data fusion in many domains mainly for two reasons. First, kernels can preserve the individual properties of data easily. Different types of kernels can be applied based on the properties of a dataset. Second, merging kernels obtained from individual datasets is easier than merging decisions in late integration (refer to the review paper (Gönen and Alpaydın 2011) for more theoretical description of kernel fusion methodologies). Followed by the seminal work of Palvidis et al. (Pavlidis, Weston et al. 2001), this idea of kernel based intermediate integration was used by Daemen et. al. (Daemen, Gevaert et al. 2007) in this context of clinicogenomic integration for classifying metastasis vs. relapse free survival of breast cancer. In particular, two normalized linear kernels were developed for both clinical and gene expression data and then, those kernels were fused using a weight before applying the final predictive model. One advantage with such kernel based integration is that the weights corresponding to an individual dataset can denote the relative contribution towards the final prediction. However, choosing an appropriate kernel for a particular dataset is not trivial. Moreover, kernels are not easily interpretable so that they can be used as biomarkers.

Graph based techniques can provide more interpretable models for intermediate integration. In an similar effort to develop such techniques, Gevaert et. al. (Gevaert, Smet et al. 2006) used Bayesian network as the intermediate representation. A Bayesian network can represent the dependency among the variables by a directed acyclic graph (DAG) in a probabilistic manner. In brief, there are two independent stages in Bayesian modeling: learning the structure of the DAG and learning the parameters of the probability distribution. The authors attempted three types of Bayesian integration-early, late and partial integration using the two independent steps of Bayesian learning. The partial integration is conceptually similar to intermediate integration. For example, first, structure learning is performed on both datasets separately (using heuristic model search algorithm K2 (Cooper and Herskovits 1992)) and then, these two structures are merged through the outcome variable which is the only common variable in the two datasets. In a second step, the Bayesian parameter estimation of the model (learning of conditional probability tables) is performed using a Dirichlet distribution. Finally, the factors within the Markov blankets of the outcome variable are defined as the biomarkers. Although such graph-based intermediate integration provides more interpretable models, merging the structures (DAGs) obtained from each datasets is not as straightforward as fusing the kernels. In both studies, intermediate and partial integration showed better performance than early and late integration.

Alternatively, many statistical approaches including canonical correlation analysis (CCA), independent component analysis (ICA), and partial least square regression (PLS) try to find latent components in each datasets, which can be also treated as a sort of intermediate representation. However, the goal of these studies is very different. In particular, these studies want to assess the relationships among the obtained components rather than building classification models using them.

Advantage: It can preserve the individual properties of a dataset. Moreover, inter-dataset relationships like correlation and redundancy can also be taken into account during final model developments, although it depends on many issues like choice of kernel and how such relationships are preserved during kernel fusion.

Disadvantage: Finding appropriate intermediate representations that are interpretable and easily fusible at the same time is difficult. Moreover, finding interactions and causal relationships across datasets is difficult due to the transformation of the original feature space.

3.2 Stage of dimensionality reduction:

Clinicogenomic integrative models have to be aware of the disparate dimensionalities of the clinical and genomic datasets. Otherwise, low-dimensional clinical variables will be lost among the thousands of genomic variables (Boulesteix, Porzelius et al. 2008). The clinicogenomic studies can also be categorized. We categorize existing clinicogenomic studies into two categories based on how they handle this issue, each of which has its own assumptions, advantages and disadvantages.

3.2.1 Two-step methods

The easiest way to handle the disparate dimensionalities of individual datasets is to first perform dimensionality reductions for each dataset separately and then, build predictive models on them in a second step. In the context of clinicogenomic integration, dimension reduction techniques are applied solely on the genomic dataset assuming that clinical variables are already low dimensional. Most of the techniques select topmost discriminative genomic features, while others methods combine those features into a combined score for future model development. In the second step, the selected genomic variables are merged with the clinical variables to build prognosis model on the combined dataset.

Advantages: The two-steps models are very flexible. Any types of dimensionality reduction technique and any predictive modeling techniques can be incorporated in building the clinicogenomic model.

Disadvantages: There are few disadvantages of the two-step methods. First, determining the appropriate number of genomic features in the first step is hard. The number of features may impact the comparison between the additive performances of clinical and genomic variables. For example, if too many features are selected from genomic data, it may overfit the clinicogenomic model in the second phase. On the other hand, if too few genomic factors are retained, then the predictive capability of the genomic factor can be underestimated. This overfitting issue is even more serious if the dimensionality reduction techniques take response variables into account in the first step. In this scenario, the genomic features fed into the second stage will have strong prediction power for the response variable. Hence, comparing those genomic features with the clinical variable is not completely fair (Boulesteix, Porzelius et al. 2008). Second, performing dimensionality reduction only on genomic data cannot account for the relationship existing between the two datasets. For example, even the right number of genomic variables selected in the first step may be redundant in the second step for model development given the clinical variables used. Moreover, the subtle contributions of many genes to prediction can be missed by the dominant genomic features that are correlated with the clinical variables (Boulesteix and Hothorn 2010). This is especially important when the goal is to assess the additional power of genomic data over clinical variables. This is described in more details in Section 0.

3.2.2 Combined clinicogenomic models

The second type of approach merges the two steps of dimensionality reduction and model development into a single step by leveraging regularization based statistical models with possible modifications. Regularized models can increase the generalization power of predictive model by preferring less complex model and thus are very effective for reducing the possible overfitting problem for high-dimensional data such as gene expression. In general, regularized techniques introduce an extra penalty term for the model complexity ($P_\lambda(\beta)$) in addition to the original loss function ($L(\beta|X)$) of the objective function as shown below.

$$\min_{\beta} L(\beta|X) + P_\lambda(\beta) \qquad \text{Equation 3-1}$$

Here, X is the clinicogenomic dataset and β is the co-efficient that represents the corresponding weight of each of the variables present in X and λ is the regularization parameter that controls the tradeoff between the loss function and model complexity. The most popular regularization approaches used in statistical learning are L_2 (ridge(Hoerl and Kennard 1970)) and L_1 (lasso(Tibshirani 1996)) regularization, which impose penalty as the square ($P_\lambda(\beta)$= $\lambda \sum_{i=1}^{p} \beta_i^2$) and absolute value ($P_\lambda(\beta)$= $\lambda \sum_{i=1}^{p} |\beta_i|$) of the regression coefficients in Equation 3-1, respectively. Moreover, L_1 penalization shrinks most of the coefficients of the regression model to zero and hence, it is widely used to perform feature selection simultaneously with model development. However, the disparate dimensionalities of clinical and genomic datasets pose new challenges to the generic regularization problem. Several modifications have been proposed to impose different penalty structures for different datasets and discussed in more details in section 4.

Advantage: The main advantage of the one-step models is that they can take the redundancy present between genomic and clinical datasets implicitly, since both datasets are considered together during model development. This property makes the single-step approach most suitable for assessing additional predictive performances of genomic features over the clinical variables (Boulesteix and Hothorn 2010). Moreover, most of the co-efficients of the sparse regularized model are zeros with few non-zero entries which precludes the explicit variable selection step. So, the number of genomic features to retain for model development is not required to be specified upfront.

Disadvantage: Each of the regularized models has their own model assumptions and requires learning several parameters. This sometimes yields to higher computational complexity. Moreover, the regression based models are mostly applicable to building predictive models. Finding inter-dataset relationships like correlation is hard using these models.

3.3 Full-space vs. Subspace modeling:

Most of the clinicogenomic models use the full space model development techniques, i.e., the bio-signatures (either gene or clinical variable) were generated based on how well they can discriminate *all* patients from the control population. However, due to the disease heterogeneity, the same set of clinical and genetic predictor may not be the causative/putative biomarker for all patients. Some factors may have more effect in a particular group of patient while the same factors may have less effect in the other group of patient (Ulitsky, Karp et al. 2008), (Fang, Kuang et al. 2010), as shown in Figure 10. This creates the need for developing techniques that are able to find not only different types of biomarkers, but also the subgroups of patients or healthy population associated with each of those particular groups of markers.

The easiest way to find the subgroup of samples associated with a set of biomarkers is to design a two-step study, where the subgroups of samples are searched in a later stage after biomarker discovery in the first phase. In one such study, Schwarz et. al. 2009 (Schwarz, Leweke et al. 2009) applied a generic network based two-step framework initially proposed by Barabasi et al. (Goh, Cusick et al. 2007) to find different subgroups of schizophrenia patients. In particular, they built a two layer bi-partite graph representing all biomarkers in one partition, and all patients in the other partition, with an edge across the two layers representing the association between them [Figure 5]. In a later phase, Markov chain clustering (Van Dongen 2000) was used to produce more homogeneous network modules containing patients with similar characteristics based on clinical state, pathological tests, brain images, and molecular information. Their network discovered one cluster containing almost a third of the Schizophrenia patients with common abnormality in serum primary fatty acid which was further validated for two psychiatric disease subtypes: naive schizophrenia and affected disorder patients. On the other hand, frequent rule mining algorithm (Agrawal, Imieli ski et al. 1993) can find patterns (e.g., the blocks A-E represented in Figure 10) representing an association between clinicogenomic factors and patient subgroups together in a single step. Berlingerio et. al., (Berlingerio, Bonchi et al. 2009) used this technique along with a postprocessing step to remove the non-discriminative patterns like (block C of Figure 10) for discovering the demographic, pathological (e.g., hepatic cirrhosis) and genomic factors (e.g., Human Leukocyte Antigens (HLA) sites) responsible for the allograft rejection of liver transplant.

Both of the previous two subspace models find subgroups of patients and the corresponding biomarkers, where biomarkers can be both clinical and genomic markers. Alternatively, some integrative studies aim to leverage the complementary strengths of the two datasets by looking at the distinct patient subgroups that the two types of markers can effectively classify. For example, clinical variables can be good at classifying a particular group of patients who cannot be classified well by genomic features and vice versa. In the context of clinico-genomic studies, preference was given to clinical data. For example, Wang et. al. (Wang, Ooi et al. 2007) first selected that subgroup of patients who cannot be well explained by the current clinical variables such as Cirrhosis and Vascular invasion commonly used for assessing the recurrence of the human hepato-cellular carcinoma (HCC) after primary treatment. Afterward, gene expression data was considered only for those subgroups for building classification models like SVM, SLD, and KNN (k=3) (Tan, Steinbach et al. 2006, Hastie, Tibshirani et al. 2009). In another recent study, Obulkasim et. al. 2011 (Obulkasim, Meijer et al. 2011) performed a more systematic study to determine automatically which samples will benefit the most by including molecular data into clinical data using step-wise classification model. At first, they built two classifiers separately on clinical and molecular training data. Second, they determined the subgroup of test samples (using a re-classification score) that either lie on the decision boundary of the classifier built upon clinical variables (assuming that clinical variables are inexpensive and well-validated) or have a chance to improve the classification accuracy if molecular data is included. In particular, they project a test sample into the clinical space of training data and then, estimate the re-classification score based on how many training samples were correctly and wrongly classified in that local neighborhood. Finally, the samples with low score were reclassified using molecular data. In a different study, Paoli et al (Paoli, Jurman et al. 2008) developed a semi-supervised approach which used the clinical and epidemiological variables to validate the coherent subgroup of patients who shared similar types of prognostic profiles defined by gene expression data. All these models can be interpreted as rules where each rule contains clinical variable as their first predictor.

Advantage and disadvantage: As mentioned earlier, the biggest advantage of subspace analysis is that it can discover patterns which are only associated with a particular group of patients. This is extremely useful for finding different types of biomarkers for heterogeneous diseases. Another big advantage of many subspace analysis techniques like association pattern mining or network based approach described here is that they are non-parametric model which can capture non-linear interaction easily. This may be extremely useful for integrating heterogeneous types of data where same kind of model assumption may not hold for all data types. Another big advantage of these approaches is that they can also be used for hypothesis discovery rather than hypothesis validation. So, they have the potential to discover novel causal factors for inferring new knowledge, especially minor causal factors that are represented in very few samples and thus, overlooked by full-space models(Fang, Pandey et al. 2010). Nonetheless, the observed patterns require more robust validation both statistically by considering the random association, and clinically by considering external cohort/test datasets before considering the patterns and modules as potential factors. Moreover, a lot of spurious patterns and modules are often discovered which are difficult to interpret.

4 Different goals of integrative studies:

In the previous section, we described the methodological differences between several integration methods based on how they address the generic challenges of data fusion. Moreover, the clinicogenomic integration can also be categorized based on the goals that they want to achieve through using those models. More or less, the overall goal of clinicogenomic studies can be divided into three broad categories from medical perspective. Some studies aim at achieving more than one clinical goal in a single study either implicitly or explicitly.

4.1 Improving the prognostic power only

Predictive clinicogenomic models aim at improving the clinical prediction of diseases by integrating clinical and genomic datasets. Thus, the main research question addressed by this type of clinicogenomic model is whether the datasets contain complementary information. To assess the improvement of prognosis power, the combined clinicogenomic method is compared with the models built on either clinical or genomic data independently. We will first describe the two-step approach which performs explicit dimensionality reduction followed by the creation of combined single-step predictive models.

Two-step models: The choice of the particular predictive model differs based on the clinical endpoints of the disease, i.e., whether the target variable is discrete or continuous. If the response variable is continuous, such as survival of patients after a particular therapeutic treatment or the development of metastasis after surgery, then the regression based methods are deployed for model development. For example, the Cox proportional hazard model estimates the lifetime (survival or failure) of an event associated with the covariates using two parameters: a hazard function describing the changes of hazard (risk) over time at the baseline level of covariates and the co-efficients describing the effect of each variable on survival. In one such clinico-genomic study, Lexin Li (Li 2006) used the Cox model for predicting the survival of the patients with diffuse large-B-Cell lymphoma (DLBCL) after chemotherapy. In addition to the genomic features (selected by a supervised dimensionality reduction (Li 1991)), they included a well-established clinical factor called international prognosis index (IPI) (Shipp, Harrington et al. 1993), which combines different clinical factors of DLBCL.

Classification techniques are used to build clinicogenomic models when the output variable has discrete categories. This includes mostly binary two-class variables, e.g., diseased vs. healthy group, successful vs. unsuccessful treatment, recurrent vs. non-recurrent, survival vs. death after certain time point, and metastasis vs. relapse free outcome. Among the wide-variety of classification schemes, discriminant models, which aim at learning a discriminative function to separate the two classes, are widely used. For example, Sun et al. (Sun, Goodison et al. 2007) used linear discriminant analysis (LDA) (Bishop and SpringerLink 2006) for combining the current clinical guidelines for breast cancer prognosis such as St. Gallen (Goldhirsch, Coates et al. 2006), (Goldhirsch, Wood et al. 2003) and NIH (Eifel, Axelson et al. 2001) with genomic information to predict the survival of breast cancer. Another popular discriminant model is logistic regression, since it can provide the probability of the outcome event in addition to learning a linear decision boundary, and thus can model clinical uncertainty. Most of the clinicogenomic studies (Stephenson, Smith et al. 2005, Beane, Sebastiani et al. 2008) use a stepwise logistic regression model where each predictive variable is added successively in the model until the optimal model is achieved. In one such model, Beane et al. (Beane, Sebastiani et al. 2008) combined the gene expression profiles of lung epithelial cells of potential lung cancer patients using bronchoscopy (Spira, Beane et al. 2007) with the clinical and demographic data to make better diagnostic decisions. Similarly, Stephenson et. al. (Stephenson, Smith et al. 2005) used step-wise logistic regression to predict the recurrence of prostate cancer after a radical prostatectomy (RP) using a well-established clinical marker called *nomogram* (Harrell Jr, Califf et al. 1982, Partin, Mohler et al. 1995, Kattan, Wheeler et al. 1999, Blute, Bergstralh et al. 2001, Graefen, Karakiewicz et al. 2002) that includes diagnostic variables such as PSA level, Gleason grade, margin status, and pathological stage along with gene expression data. For avoiding model over-fitting, a goodness of fit measure like Akaike's information criteria (Akaike 1974) is used to select the optimal model. Another popular discriminant model is the support vector machine (SVM) (Vapnik 2000), which maximizes the separation between the two classes (margin) to achieve better generalization power in unseen datasets. Li et al, 2005 (Hoeting, Madigan et al.) applied SVM to predict the survival of advanced-stage ovarian cancer after platinum-based Chemotherapy. SVM can also learn a non-linear decision boundary using the kernel trick (Appendix B) which was used for developing an intermediate integration described earlier (Section 3.1).

Other types of non-linear models have also been applied for the integrative purpose. For example, tree based methods (Hastie, Tibshirani et al. 2009) are very popular, since they can be easily represented as classification rules which are more interpretable to clinicians and can be tested for inferring new domain knowledge. These methods are based on recursive partitioning of all available samples into more homogeneous subgroups with respect to the binary class variable. One early attempt to use tree based method was conduccted by Pittman et al. 2004 (Nevins, Huang et al. 2003, Pittman, Huang et al. 2004) to integrate genetic data with clinical variables for enhancing the prognostic power of breast cancer patients relative to long-term recurrence. Similarly, Clarke et al, 2008 (Clarke and West 2008) developed a clinicogenomic model for the survival prediction of ovarian cancer. One problem with tree based methods is that there is no single optimal tree because they are built using heuristic search criteria. To circumvent this problem, all these clinicogenomic studies used ensemble learning (Hothorn, Buhlmann et al. 2005),(Kittler 1998) and model averaging (Oliver and Hand 1995, Raftery, Madigan et al. 1997, Hoeting, Madigan et al. 1999) techniques to generate a forest of trees and then, estimate the final prediction by taking the weighted average of the individual predictions of each tree. Such techniques not only boost the predictive performances by combining many weak learners (trees), but also provide a confidence interval for the prediction estimated from the individual models. This property is extremely useful in the context of an integrative clinicogenomic study for capturing the clinical uncertainties (Kelley and Ideker 2005, Calnan 2008) arising from different clinical processes such as variability of tissue processing, hybridization measures, small sample size, and sample selection (Nevins, Huang et al. 2003, Pittman, Huang et al. 2004). Also, such model uncertainty may capture potential conflicting predictions either within or between the clinical and genomic factors, which can be very important for complex heterogeneous diseases. Similarly, mixture of expert (ME) is another non-linear method that combines several expert trees using a convex weighted sum of all the outputs produced by them. However, each expert can be trained on different partitions of the input data with possible overlaps among them (soft split) as opposed to hard split of the data used by CART. Cao et al. (Lê Cao, Meugnier et al. 2010) applied ME method for integrating categorical clinical variables directly with continuous-valued gene expression data without any discretization. Furthermore, ME provided better result than random forest based approached used by (Boulesteix, Porzelius et al. 2008).

Single-step sparse models without explicit dimensionality reduction: Some clinicogenomic studies leverage the strength of sparse modeling technique to perform model development and feature selection in a single step by considering clinical and genomic data simultaneously. For example, Ma et al, 2007 (Ma and Huang 2007) extended one such iterative boosting approach called *Threshold Gradient Directed Regularization* (TGDR (Friedman and Popescu 2004)) into a more generalized framework (Cov-TGDR) for two generalized linear models: logistic regression and the Cox survival model. Cov-TGDR iteratively optimized the gradient of negative log-likelihood considering as the loss function ($L(\beta|X)$ in Equation 3-1). Moreover, in each iteration the component-wise gradient was updated only for only a few variables controlled by a regularization parameter λ. Thus, the components with lower gradient values are not updated in each iteration and these results in a sparse representation of the solution (β). Moreover, variable selection was performed separately for the two datasets to respect their individual properties of the data using two parameters λ_1 and λ_2 for the two datasets in Equation 3-1. Finally, this study applied the Cox proportional model for the survival of follicular lymphoma (Dave, Wright et al. 2004) and logistic regression for the binary prediction of the development metastasis of breast cancer (Van't Veer, Dai et al. 2002).

Comparative studies: van Vilet et al. (van Vliet, Horlings et al. 2012) performed a recent comparative study of two-step predictive models to systematically assess whether combining clinical and genomic data help improve the prediction power of breast cancer. They consider three simple classifiers such as nearest mean classifier (NMC), Naïve Bayes, Nearest neighbor, and two more complex classifiers such as SVM (similar to (Daemen, Gevaert et al. 2007)) and tree based classifier. All of these models were developed in three different stages (early, intermediate and late) along with no integration (built on clinical and genomic variables). The original tree based classifiers proposed by (Pittman, Huang et al. 2004) were modified for intermediate integration by restricting one dataset at the top node. For all these classifiers, integration improved the prediction power for breast cancer significantly, and simple classifiers performed better than complex classifiers (with NMC with OR-type late integration performing the best) which may be an effect of small sample size. Moreover, either late or intermediate strategies performs the best, which confirms the previous studies by (Gevaert, Smet et al. 2006, Daemen, Gevaert et al. 2007). Unlike the previous study by (Van't Veer, Dai et al. 2002), this study found that clinical data has slightly better information than genomic data, which they believe that is mainly because of more comprehensive clinical features such as matrix information, central fibrosis, etc. Moreover, the genomic and clinical features obtained from this study perform better than the markers found by previous four studies in different cell lines (Van't Veer, Dai et al. 2002, Chi, Wang et al. 2006, Sotiriou, Wirapati et al. 2006, Liu, Wang et al. 2007). However, they did not assess the effect of different feature selection techniques in the model development stage. Bovelstad et al. (Bovelstad, Nygard et al. 2009) provided a methodological comparison of different dimensionality reduction techniques designed for Cox regression in survival studies. They covered both two-step and one-step approaches (Figure 4) in their model development and they observed that modified ridge regression performed the best when applied to three different clinicogenomic datasets. However, they did not compare it to the Cov-TGDR methods.

4.1.1 Advantages and disadvantages of the predictive models

The main advantage of predictive models is that they are easy to develop and simple from a methodological perspective. Any model that is applicable on either clinical or genomic data can be applied directly (for two-step approaches) or with minor modifications (for regularized methods) to the combined dataset. These models build unbiased models on clinical and genomic data sets without any prior information and bias towards any of the datasets being integrated. Therefore, the predictive model can test whether the datasets being integrated are complementary in nature based on the improvement of the predictive power of the combined model over the individual models. However, the final clinicogenomic models may select a completely different set of clinical and genomic variables than those selected by independent models. Hence, comparing the predictive power of clinical and genomic features grossly in dataset level cannot assess directly how much additional power genomic features possess given the traditional clinical variables. This is an interesting question for clinicogenomic integration as described in the next section.

4.2 Assessing additive prognostic effect of clinical variables over the genomic factors

The predictive clinicogenomic models described in the previous section treats clinical and genomic datasets similarly. However, clinical variables are considered more important than genomic variables by many studies for two reasons. First, clinical variables are well-validated through independent studies unlike genomic factors. Second, clinical factors are easy to collect and currently used in the healthcare system, and thus reuse of those clinical variables will also reduce health care costs. Therefore, treating both datasets similarly may underestimate the clinical variables and overestimate the performance of the genomic variables significantly. Trutzner et al. (Truntzer, Maucort-Boulch et al. 2008) performed a systematic study to assess such optimistic use of genomic markers over the clinical variables. Using the synthetic datasets, the authors showed that the genes selected by the unbiased predictive models are less reproducible in the independent test datasets. They used both the two-step methods and the Threshold Gradient Directed Regularization (TGDR (Friedman and Popescu 2004)) described in the section 4.1, and concluded that such over-estimation of the value of genomic data increases because of the estimation of too many free-parameters for large number of genes with small samples. The two-step methods containing separate supervised dimensionality reduction step are even more prone to over-estimation (Section 3.2.1). For such two-step methods, Tibshirani et al. (Tibshirani, Efron et al. 2002) performed some seminal works and proposed the pre-validation framework to compare the genomic markers to clinical markers more rigorously. In particular, they suggested that the genes should be selected by a separate cross-validation framework rather than the same cross-validation framework used for assessing the predictive performance of the final model (more detail in the validation section). In contrast, for models built for one-step combined clinicogenomic study, it is less difficult to remove such over-estimation.

In addition to categorization based on how dimensionality reduction is performed, clinicgenomic studies can be further categorized into two groups based on how the additive power is assessed. One type of study builds clinicogenomic models that are biased to the clinical markers by including the clinical variables (or clinical index built thereof) as a mandatory variable in the model development phase. The second type of study focuses directly assesses the additional power of the genomic data given the clinical variables using a hypothesis testing framework. Strictly speaking, they answer the question of ‘Do genomic variables boost the performance of models given the clinical variables?’ in compared to the null-hypothesis of ‘no additional value’.

Developing clinicogenomic models biased towards clinical variables: Using the ‘pre-validation’ framework provided by Tibshirani et al., Boulesteix et al. 2008 developed a two-step clinicogenomic model which can assess the additional predictive power of genomic data using two separate cross-validation loops, one for each of the two-steps. The first cross-validation was used to reduce the genomic features to a few unbiased pre-validated components (Tibshirani, Efron et al. 2002) using the supervised partial least square (PLS) method (Wold 1985). Second, they built a random forest (Breiman 2001) which first selected all clinical variables as mandatory variables and then added PLS genomic components one by one, as long as the predictive power improved as assessed by the out-of-bag (OOB) error (Breiman 2001) using a bootstrapping strategy. Therefore, the additional performance was assessed by the number of genomic components selected automatically by the predictive model in addition to the clinical features. However, as discussed in Section 4.2, two-step methods are only partially successful in removing potential redundancy that can be present between the clinical and genomic features. For example, in the previous study, some of the PLS components that have marginal predictive power are non-redundant compared to the clinical variables may be missed in the first phase. Alternatively, a single-step sparse Cox model called *CoxBoost* has been proposed by Binder et al. 2008 (Binder and Schumacher 2008) to assess the additional power of genomic data for survival study using a component-wise offset based boosting approach (Tutz and Binder 2007). In particular, they optimize the log-likelihood of the model via component-wise gradient boosting updates using a Neuton-Raphson method. Moreover, all the clinical variables were included in the model as the mandatory variables using a customized diagonal penalty matrix with ‘zero’ entries (in the second term of Equation 3-1) while feature selection was performed on the genomic features using ‘one’ entries in the penalty matrix to assess the additional predictive power. In a similar study, Kammer et al. (Kammers, Lang et al. 2011) used other types of sparse models such as L_1 and L_2 regularization techniques only on gene expression data and included clinical variables as mandatory variables into the model.

Besides the fullspace based models described so far, several subspace-based models also have been developed where genomic variables are only included for the subgroup of patient that cannot be predicted by clinical variables. An initial attempt to develop subspace based models stratified the population based on the clinical data and then, included genomic data for each subgroup to improve its prediction power. For example, estrogen receptor status was used by (Wang, Klijn et al. 2005, Teschendorff, Naderi et al. 2006, Teschendorff, Miremadi et al. 2007), while Dai et al. (Dai, van't Veer et al. 2005) used other traditional clinical variables, such as age, tumor grade and tumor status for stratifying breast cancer population. Similarly, Wang et. al. (Wang, Ooi et al. 2007) used cirrhosis and vascular invasion which are commonly used for assessing the recurrence of the human hepato-cellular carcinoma (HCC) after primary treatment. These studies are similar to decision tree rules where the topmost nodes are restricted to be from clinical variables. On the other hand, (van Vliet, Horlings et al. 2012) explicitly developed one such hybrid tree based on intermediate approach where the prediction obtained from a classifier was used for the topmost node. In another recent study, Obulkasim et. al. 2011 (Obulkasim, Meijer et al. 2011) performed a more systematic study to determine automatically which samples will benefit the most by including molecular data into clinical data using step-wise classification model. First, they build two classifiers separately on clinical and molecular training data. Second, they determine the subgroup of test samples (using a re-classification score) that either lie in the decision boundary of the classifier built upon clinical variables (assuming that clinical variables are inexpensive and well-validated) or has chance to improve the classification accuracy if molecular data is included. In particular, they project a test sample into the clinical space of training data and then, estimate the re-classification score based on how many training samples were correctly and wrongly classified in that local neighborhood. Finally, the samples with low score were reclassified using molecular data. The problem with such subspace based methods is that there may not be enough samples associated with each subgroup, so building classification models in each subgroup may be difficult and statistically insignificant.

Hypothesis testing frameworks: All the biased clinicogenomic models discussed so far assess the additional power of genomic features indirectly using how many genomic features are included in the model. However, some of the selected components may be statistically insignificant. The more effective way to address this issue is to assess the additive performance of genes in a hypothesis testing framework. In a seminal study, Tibshirani et al. (Tibshirani, Efron et al. 2002) first summarized the genomic variables into a single unbiased genomic score using the pre-validation framework (LASSO internal model). In a second step, a hypothesis testing framework was designed based on linear regression model (or any GLM) built on the clinical variables and the pre-validated genomic (PVG) markers used as 'pseudo-predictors'. In particular, the added predictive value was assessed by whether the regression coefficients of the genomic marker was statistically significant, i.e. $\beta_{PVG} > 0$ compared to the null-hypothesis $\beta_{PVG}=0$ using t-test or z-tests. In a later study (Höfling and Tibshirani 2008), they showed that this test was biased because of the violation of the i.i.d. assumption by the sampling procedure used in the PVG framework uses regression. Alternatively, they proposed a random permutation based empirical p-value estimation. In both case, it was shown that pre-validated genomic score was less significant than the genomic score without pre-validation using a landmark breast cancer study (Van't Veer, Dai et al. 2002) which actually over-estimated the performance of gene expression data. However, any two-step approach cannot remove potential redundancy between the clinical and genomic data completely (Section 3.2). For example, if clinical and genomic markers are correlated, then both types of markers will have significant coefficients by the above approach.

A more rigorous hypothesis testing framework has been proposed by Boulesteix et al. (Boulesteix and Hothorn 2010) considering both types of datasets simultaneously in a similar way as CoxBoost (Binder and Schumacher 2008) to remove any types of redundancy between the two types of variables completely. The main idea of the method was to include not only the clinical variables, but also the contribution of those clinical variables as the mandatory variable in the clinicogenomic model, so that genomic variables cannot influence the clinical contribution. More specifically, this method first fits a generalized linear model on the clinical variables only, and then the clinical predictor is used in the final combined clinicogenomic model built by least-square boosting strategy (Friedman, Hastie et al. 2000) as a fixed offset such that its co-efficient is not changed during the iterative learning process. Thus, the genomic features cannot affect the contribution of the clinical features in the final model, unlike the CoxBoost and Pre-validation methods. Finally, the likelihood of the boosting method was tested for the statistical significance by randomly permuting the genomic variables to estimate the additive performance similar to (Höfling and Tibshirani 2008). Although this approach did not perform any feature selection for genomic features similar to CoxBoost, it can be easily generalized to a regularization based framework, as argued in the later study (Oelker and Boulesteix). In this later study, they also compared both pre-validation testing and Globalboosttest by generating several synthetic datasets with different amounts of correlation between clinical and genomic markers. As expected, if the informative genes are perfectly correlated with the clinical variables, Globalboosttest is more conservative in selecting genomic features (p-values uniformly distributed in [0, 1]) than pre-validation. Note that the pre-validation framework only removes the bias associated with genomic variables, but can compare the two datasets in a more generic fashion. In contrast, Globalboosttest is completely biased to clinical variables with the sole purpose of rigorously testing the additional power of genomic marker, but the opposite properties of the two datasets cannot be tested.

Incorporating prior knowledge: One issue with including gene expression data directly into the model is that the selected genes do not necessarily yield to biologically interpretable pathways. Moreover, each of the genes belonging in a pathway may have weak association and thus missed by the model, but their aggregate association may be large. Testing the association of the pathways with disease directly, rather than in a post-processing stage has become popular to aid clinical interpretability (Subramanian, Tamayo et al. 2005). Kammer et al. (Kammers, Lang et al. 2011) recently also used gene ontologies (GO) for grouping the genes and then the combined effect of each GO group (assessed by the first principal component) as the predictor in a Cox survival model. However, some GO groups are very generic and only part of a GO process can be activated in a particular disease due to disease heterogeneity. Alternatively, the author also further clustered the genes belonging to each GO group into several subgroups before including them into the model. From a methods perspective, they followed the combined one-step model development where both L_1 and L_2 penalization scheme (Equation 3-1) were used for handling high dimensional genomic data by including clinical variables as mandatory variables. All the three types of genomic data, i.e., the original gene expressions, GO groups and pre-clusters of GO groups when combined with the clinical variables provided similar performance assessed by p-value of the final prognostic model and Brier score. Since the pre-clustering technique is unsupervised here and guided by GO, no pre-validation like framework was necessary to reduce the bias of the genomic data as well.

Advantages and disadvantages: The main advantage of the pre-validation based framework is that it can compare the genomic and clinical features more directly by removing any sort of redundancy among them and thus can assess the additional predictive power of the genomic features in an unbiased manner. However, the pre-validation based framework combines the genomic features into one or more newly developed features, which make the interpretation of the final model difficult for biomarker discovery. Another problem with such models is that they assume that clinical variables are important and thus the predictive models should be biased towards clinical variables. However, this assumption may not be true in the future as genomic data become more easily available and are validated in multiple independent studies. Moreover, sometimes the clinical variables, such as pathological and behavioral effects, can be the downstream effect of causal genomic features. In that case, genomic features may not provide additional predictive power over clinical variables. However, knowing such relationships among different types of markers can be useful knowledge. Neither the original predictive models nor these models unbiased genomic scores aim to assess the relationships present among different types of data.

4.3 Assessing relationship between the clinical and genomic studies

The previous two types of studies mainly aimed at building predictive models by integrating clinical and genomic data. It is also important to understand novel knowledge about diseases by looking at the relationships present among different datasets, e.g., correlation, interaction and causality. Each type of relationship may reveal novel insights about the complexity of human disease. For example, if some kind of causal relationship can be inferred between the molecular and clinical factors or vice versa, drugs can be designed in a better way to target the original causal factor or preventive health care can be designed in a smarter way based on the understanding of these two factors. Finding such relationships among these datasets and with genetic and genomic factors has become very popular recently. The most prominent application of such approaches is in neuroscience domain, where abundant clinical and pathological data are collected using MRI technology that measures various brain activities. Examples include fMRI, DTI, and sMRI data, which provide information about the functional and structural connections among brain regions, and volumetric information of brain, respectively.

Several multivariate statistical models have been extended for finding inter-dataset relationships among clinical and genomic variables, where inter-dataset relationships are measured by an association measure such as correlation. In addition to the inter-dataset relationships, there can also be intra-dataset relationships present in the datasets. For example, the nearby locations in the brain can behave similarly leading to spatio-temporal autocorrelation. To reduce the redundancy present with a single dataset, several blind source separation based techniques such as principal component analysis (PCA (Jolliffe 2002, Park and Hastie 2007)) and more generic independent component analysis (ICA) (Hoerl and Kennard 1970) have been found useful. In brief, the blind source separation techniques generate two matrices from a dataset: the modulation profile and the component maps, where component maps represent the sources and the modulation profile denotes the association of each individual with those components. These techniques are useful for reducing both the dimensionality and spatio-temporal correlation present in several neuroscience datasets including fMRI (Biswal and Ulmer 1999), sMRI (Xu, Pearlson et al. 2009), EEG (Delorme and Makeig 2004) and others (Teschendorff, Miremadi et al. 2007, Cichocki, Zdunek et al. 2009). These original BSS methods have been extended for integrating multiple dataset with the goal of finding relationships among multiple datasets. For example, the easiest way to extend the ICA based framework for multiple modalities is to combine the two datasets being integrated into an augmented matrix, as in the early integration technique described earlier and then, perform ICA on that augmented matrix to find the common modulation profile and a single component map. Note that each component represents the features from both datasets being integrated. This technique called joint ICA (Calhoun, Adali et al. 2006, Gönen and Alpaydın 2011), has a very strict assumption that each sample is modulated in the same amount in two datasets, which may not be true for all cases.

Alternatively, several statistical methods like canonical correlation analysis (CCA) and parallel ICA (pICA) provide a more natural framework for data integration where the relationship between different components found from multiple datasets is defined as the inter-subject variabilities. In general, these models decompose each dataset into two components: a modulation profile and a component map using any blind source separation techniques such that the components across the datasets are related somehow over the samples. For example, CCA (Correa, Adali et al. 2010) wants to find a linear transformation of the data (modulation profile) for each dataset such that they have maximum correlation after transformation. Therefore, the inter-dataset variability is measured by correlation(Bay and Pazzani 2001, Correa, Adali et al. 2010). Multi-set CCA (Dai, van't Veer et al. 2005) and a generalized CCA which can integrate more than two datasets, has also recently been applied for integrating fMRI, EEG and sMRI datasets (Lê Cao, Martin et al. 2009). One major problem with applying all these multi-variate models is the original overfitting problem described earlier in Section 2. Another issue is that the individual components are not directly interpretable due to the linear combination over the selected subset of features. To circumvent these issues, Cao et al. (Obulkasim, Meijer et al. 2011, Westra, Dey et al. 2011) recently have proposed several alternative optimization formulations of the original CCA. More specifically, they introduced both the L_2 and L_1 norm in the penalty formulation for reducing both model overfitting and performing variable selection at the same time, which is similar to the elastic net technique(Tibshirani 1996). However, most of these studies cannot take the class label into account while finding the canonical components, and thus generate many non-discriminative components that are pruned in a later stage. Recently, Sun et al. (Truntzer, Maucort-Boulch et al. 2008) proposed a discriminative CCA (DCCA) technique, which can take the class labels into account while finding canonical components.

Parallel ICA (Chi, Wang et al. 2006, van Vliet, Horlings et al. 2012) is an alternative technique, but the components are based on the original ICA instead of the linear transformation procedure of CCA. Each of these multi-variate models has their own assumptions (Correa, Adali et al. 2010). It has been also shown that CCA has fewer model assumptions than ICA based techniques, since CCA is based on the second-order statistics (Correa, Adali et al. 2010). Recently, an effort has been made to combine these two techniques to minimize such assumptions (Liu, Wang et al. 2007, Sui, Adali et al. 2010).

5 Validation

In this section, we discuss the validation procedures of the clinicogenomic models described so far. Since the main goal of all these clinicogenomic models is to improve the prognostic power of disease, they compare the combined clinicogenomic model with the models built on either genomic data or clinical data alone. We will first discuss several performance metrics used for this purpose. Then we will discuss different validation techniques to assess the effectiveness of obtained results from clinicogenomic models.

Performance metrics: The most common metrics for performance measurement of the binary classification based models (Beane, Sebastiani et al. 2008), (Sun, Goodison et al. 2007) are accuracy, precision, recall and area under the ROC curve (Tan, Steinbach et al. 2006). On the other hand, the studies that want to predict continuous outcome variable such as survival time and disease progression-free probability (PFP) use different metrics e.g., c-index, to assess (Stephenson, Smith et al. 2005) how well the model discriminates between patients with different survival probabilities. C-index measures the concordance between the predicted and observed responses (Harrell Jr, Califf et al. 1982) in a scale between 0-1. Another popular measure used by (Li 2006) is the time dependent area under the curve (AUC) defined by (Heagerty, Lumley et al. 2000). On the other hand, instead of using cross-validation, Binder et al. (Binder and Schumacher 2008) used a bootstrap sampling strategy, as in (Schumacher, Binder et al. 2007), for performance evaluation using the Brier score (Gerds and Schumacher 2006). Some studies (Beane, Sebastiani et al. 2008),(Tibshirani, Efron et al. 2002) also used the co-efficient of the genomic and clinical markers to estimate their relative contribution towards the predictive model. However, the performance gain can be obtained by random chance as a mere data artifact, thus yielding overoptimistic results unless they are validated for statistical significance or repeatedly observed in multiple datasets (Boulesteix 2010). Permutation based techniques have also been used by many studies to get the statistical significance of the observed result. For example, (Höfling and Tibshirani 2008) randomly permutated a genomic marker X to get the statistical significance of the observed coefficient of genomic marker. Similarly, (Stephenson, Smith et al. 2005) permutated the class label to get the statistical significance of the classification accuracy of the predictive model. Some studies (Li, Chen et al. 2005, Beane, Sebastiani et al. 2008) also used standard hypothesis tests-Wilcoxon test, t-test, and z-score-to get the statistical significance of the improvement in performance of combined model over the individual models. Beside all these measures, the Kaplan-Meier curve (Kaplan and Meier 1958) is a popular visualization technique to visualize the survival probabilities of different groups of population along the progression of time. All clinicogenomic survival studies used this technique to visualize the prognostic separations of subpopulations defined by the final model.

Besides estimating the performance of the predictive model empirically using the above mentioned metrics, some clinicogenomic studies validated their obtained results from a domain perspective as well. Some studies wanted to investigate which groups of patients benefitted the most by the integration of clinical and genomic markers. For example, Stephenson et al. observed that their clinicogenomic model can significantly improve the prediction of a sub-sample (~30% of the whole prostate cancer dataset) where the prediction of well-established clinical monogram is in middle range (7-year PFP, 30-70%). On the other hand, (Beane, Sebastiani et al. 2008) validated their observed combined clinicogenomic model by three expert pulmonary physicians. Some studies, e.g., (Stephenson, Smith et al. 2005), tried to find biological information about the obtained predictors in previous literature. The breast cancer studies (Pittman, Huang et al. 2004, Clarke and West 2008) found that the important clinical factors (lymph node status and estrogen receptor (ER) status) and metagenes selected by the topmost trees were well recognized in clinical practice and had been validated through previous study. For example, all of these studies identified some of the metagenes that are related to estrogen pathways or growing signal pathways, or are correlated with the ER status. The breast cancer study by (Sun, Goodison et al. 2007) compared their obtained model with the 70-gene signature built by Veer et. al. (Van't Veer, Dai et al. 2002). Ma et al. (Ma and Huang 2007) also confirmed the obtained significant genes from previous studies.

Validation procedures for predictive models: The ideal technique for testing the obtained model is to use an external validation dataset that is collected independently (Boulesteix and Sauerbrei 2011) of the training dataset on which the model was built. For example, Beane et al. 2008 compared the performances of clinicogenomic model on independent test data sets that did not have a definite diagnosis following bronchoscopy as a part of diagnosis for lung carcinoma. However, in most of the practical cases, data is scarce and expensive to collect. It is also hard to design similar experimental setups for collecting both validation data and the training data in an unbiased manner. The simplest way to solve this problem is to divide the original data into two disjoint sets: training and test data. The training data is used to develop the model while test data tries to mimic the independent validation data. For example, some clinicogenomic studies (Li 2006), (Bair, Hastie et al. 2006, Beane, Sebastiani et al. 2008) use a simple set up of random splitting of available data based on previous studies. Alternatively, such random splitting is repeated several times by some studies to avoid selection bias (Clarke and West 2008, Bovelstad, Nygard et al. 2009).

K-fold cross-validation (Kohavi 1995), (Hastie, Tibshirani et al. 2009) provides a more systematic framework by dividing the available data into K parts, where each of these K parts is considered as a test set while the rest of K-1 datasets are considered as the training set. Bootstrapping (Hastie, Tibshirani et al. 2009), which is another useful validation technique, samples the original data with replacement to estimate the variance of the result. Applying standard techniques such as cross-validation or bootstrapping for the one-step regularization based techniques is straightforward. However, applying them for two-step approaches is not straightforward because of the separate supervised dimensionality reduction step. In the simplest setting for building a two-step predictive model, the first step of dimensionality reduction is performed on the whole dataset and the second step of predictive model development is performed using two separate datasets: a training dataset for learning the model and a test dataset for assessing the performance of the observed model. However, as mentioned in (Smialowski, Frishman et al. 2010), (Simon, Radmacher et al. 2003), performing supervised dimensionality reduction on whole dataset provides biased results because of the use of test data set, on which the performance of the final predictive model is estimated in the second step. In order to get an unbiased estimate of the performance, both the supervised dimensionality reduction step and predictive model development should be performed solely on training dataset. Figure 6 describes this phenomenon. The correct set up is shown as the step A followed by the step C. More specifically, the whole two-step development process of Figure 4(a) should be done in step A of Figure 6, which is shown in detail in Figure 7. In particular, the dimensionality reduction is only performed for the genomic dataset and then the obtained features (some selected genes or some newly developed features) are combined with the clinical marker using a predictive model. Several studies ((Stephenson, Smith et al. 2005), (Pittman, Huang et al. 2004), (Li, Chen et al. 2005), (Sun, Goodison et al. 2007), (Ma and Huang 2007) using LOOCV, (Pittman, Huang et al. 2004)) are aware of the fact and used this step to assess the predictive models correctly.

Sometimes, models have a few parameters that need to be learned from the data itself. In those cases, one more inner CV is used to select the optimal parameters for the classifier for the training data obtained from the outer CV framework ((Sun, Goodison et al. 2007), (Ma and Huang 2007) and (Binder and Schumacher 2008) used 5-fold CV and (Bovelstad, Nygard et al. 2009) used 10-fold internal CV). In subsequent discussions, we will ignore this inner CV for simplicity and assume only one CV for estimating the final performance of predictive model.

Assessing additional predictive values:

The experimental design of methods with the goal of assessing additional predictive power should be performed carefully; otherwise the prognosis effect of the genomic marker may be over-estimated. For example, if supervised feature selection techniques are used in the validation procedure shown in Figure 7, another issue arises. Specifically, the genomic features are either selected or created in such a way so that they are the most discriminating features in the original genomic dataset **X**. Comparing such discriminative genomic feature with the clinical variables can lead to over-estimation of the predictive power of the genomic features. If we look at the last training phase of model development, the genomic data $\mathbf{X_{tr}}$**'** already have seen the label $\mathbf{y_{tr}}$ but the clinical features have not.

Tibshirani et al. proposed a variant of the cross-validation framework (called *pre-validation*) to remove such kind of bias toward genomic variable. In particular, they proposed one more k-fold CV for supervised dimensionality reduction even before developing predictive model using the second CV framework. The available training data ($\mathbf{X_{tr}}$) will be further divided into two sets as mentioned by (Tibshirani, Efron et al. 2002). One set will be used for the dimensionality reduction step to select and/or create genomic feature out of the original data and then other set of data will be used for building predictive model on the combined clinical and the obtained features from the previous step. The detailed steps are described below (Figure 8):

1. Divide the available training data into k separate parts.
2. The first (k-1) parts are used to learn the dimensionality reduction to select and/or create genomic feature out of the original data.
3. Afterward, the same set of selected genes or feature creation rules will be applied to the left-out k-th samples to predict the label of them.
4. Repeat the steps 2 and 3 for each k-th part to get the unbiased predictor of genomic variables for all samples.
5. Build the predictive model on the combined clinical and the pre-validated genomic features. Comparison between the clinical and genomic factors can be done as well here.

Tibshirani et al provided both theoretical and empirical evidence that the 'pre-validated' genomic score has fewer degrees of freedom (ideally one) than the non-validated version. So, this score can be treated as a 'fairer' pseudo-predictor as if it was built on an independent dataset, and hence the whole dataset can be used for model development in step 5. They also empirically showed that a pre-validated genomic factor is less significant than that of a non-pre-validated predictor when compared to the clinical variables. Although the authors used the above technique for summarizing all genomic factors into a single predictor, it can be easily generalized for selecting more than one feature as performed in (Boulesteix, Porzelius et al. 2008). In summary, the right setup for both developing predictive model and assessing additional predictive power of the genomic features is as follows:

1. To estimate the performance of a two-step predictive model along with a supervised dimensionality reduction technique, separate training and testing data are required as shown in Figure 6.
2. To compare the clinical and reduced genomic factors fairly, the supervised dimensionality reduction and comparison of genomic variables with the classical clinical predictors should be done in separate datasets or the pre-validation technique should be used to build a fairer version of the genomic predictor (Figure 8). This step is critical for assessing the additional performance gain from genomic data.

Here, in both of the steps where separate training and testing set are required several alternative techniques like repeated sampling or boosting can be used along with cross-validation. Boulesteix et al.(Boulesteix, Porzelius et al. 2008) performed these two steps using two separate cross-validation loops. The first CV was for the pre-validation of selected genes as mentioned by Tibshirani et al. (Tibshirani, Efron et al. 2002) with the second one for estimating the classification error rate of the random forest (Breiman 2001) model built on the selected gene signature component from the previous step and the clinical data.

6 Integrating homogeneous and heterogeneous datasets/multi-site integrative studies

Besides integrating datasets coming from multiple modalities/sources (integrating heterogeneous datasets), integration can also be performed to combine multiple similar types of datasets (integrating homogeneous datasets). Especially, genomic data are often criticized for the lack of reproducibility among the independent cohorts. For example, very few overlapping genes were observed between the biomarker genes of the two well-known breast cancer studies by (Van't Veer, Dai et al. 2002) and (Wang, Klijn et al. 2005) by other independent studies (Ein-Dor, Zuk et al. 2006), (Ein-Dor, Kela et al. 2005), (Chuang, Lee et al. 2007), (Naderi, Teschendorff et al. 2006). The main reasons for such poor consistency of genomic signature across studies are “small sample cohorts size, selection bias during sample inclusion and annotation, different protocols for sample preparation and data preprocessing, and heterogeneous clinical endpoints for different studies (Shedden, Taylor et al. 2008)”. Therefore, integrating multiple cohorts of same kind of patients can increase the sample size significantly and thus, is very popular to develop reproducible genomic biomarkers (Fan, Oh et al. 2006). Such multi-site integration can be performed in many ways: either keeping the most common features among all datasets (data level early integration), or by learning a more sophisticated Bayesian method to fuse information available in individual datasets (Troyanskaya, Dolinski et al. 2003), (Konstantinopoulos, Cannistra et al. 2011). Inspired by such multi-site studies, some clinicogenomic studies integrated not only multiple gene expression datasets, but also multiple clinical datasets to build multi-site universal clinicogenomic models and finally assess the improvement of prediction power of such models over that of multi-site genomic biomarkers (Table 5).

Predictive multi-site studies use some of the available independent cohorts for developing clinicogenomic model and then use the rest of the cohorts for testing the reproducibility of the predictive model. The existing multi-site clinicogenomic models proceed in two steps (Figure 9). For homogeneous integration in the first step, most of the studies take the simplistic approach of retaining only those features that are common in all of the datasets. Then, the heterogeneous integration of clinical and genomic data is performed using any of the techniques described earlier. For example, Teschendorff et al. (Teschendorff, Naderi et al. 2006) built a universal molecular prognosis marker from five publicly available gene expression datasets including their own collected gene expression data for breast cancer survival prediction. They used three cohorts for building a Cox regression based predictive model and reserved the other three as independent test sets. However, instead of using classification accuracy for validation, they used a recently developed statistical distribution based evaluation measure called *D-index* (Royston and Sauerbrei 2004), which depends only on the relative risk ordering of the test samples rather than relying on the absolute value of outcome variable. Thus the prediction power remains unchanged as long as the relative ranking of the test samples are not changed. This property makes the D-index suitable for assessing the performances over test samples coming from different cohorts with diverse characteristics. On the other hand, Shedden et al.(Shedden, Taylor et al. 2008) tried to minimize the experimental bias in multi-site studies directly by generating their own datasets from six different institutions using a uniform robust and reproducible protocol (Dobbin, Beer et al. 2005). Moreover, several different gene selection methods along with different classifiers were applied on two out of four datasets collected for predicting the survival of lung adenocarcenomas patients. Congruent with most of other clinicogenomic studies, clinical variables like cancer stage and age added some prognostic power to gene expression, especially for more heterogeneous stage-1 lung cancer patients.

In another large integrative study, Acharya et al. 2009 (Acharya, Hsu et al. 2008) integrated five breast cancer cohorts with the goal of identifying additional breast cancer subtypes, which have different underlying biological mechanisms beyond their common clinic-pathological characteristics. At first this study divided all patients into three main risk categories (low, medium and high) using a clinical guideline called Adjuvant online! (Goldhirsch, Coates et al. 2006), which uses age, tumor grade, tumor size, and lymph node status as predictors. Afterward, the study refined each of these patient groups into more homogeneous subgroups based on the clusters found from the genomic data. Moreover, they also incorporated prior knowledge by selecting genes that are involved in several pathways such as altered tumor microenvironment states, oncogenic pathway deregulation and chemotherapy response. Another contribution of the paper was that they also assessed the chemotherapy sensitivity of different clusters based on the underlying oncogenic pathway and tumor micro-environmental deregulation, which can help design better therapeutics for breast cancer patients. Although the clinicogenomic model was not compared directly with the clinical model, this study provides an indirect result that gene expression data has some additional prognostic power over the traditional clinical factors.

7 Discussion

Clinicogenomic integration has received wide attention from different communities recently, because of its great potential of integrating diverse perspectives from clinical and genomic sources to reveal complex disease mechanisms. Because of the multi-disciplinary nature of the topic, the approaches taken by all these clinicogenomic efforts are quite diverse, although the objective is same: improving the prognosis power of predictive models for complex diseases. In this article, we survey these clinicogenomic studies with emphasis on, but not restricted to the methodological perspective. We aimed at finding the existing challenges in integrating heterogeneous datasets such as clinical and genomic data, and understanding how these challenges were handled by the methods in the clinicogenomic context. This review can also be relevant for some other integrative studies, as well, where the challenges are similar to those for the integration of clinical and genomic data. For example, Hamid et al. [(Hamid, Hu et al. 2009)] proposed a theoretical framework for integrating different kind of genomic data that has some common challenges with the clinicogenomic integrative efforts. Thus, some of the integrative methods can be shared between both areas.

The main purpose of most of the clinicogenomic studies was to develop better predictive model for complex diseases through integration. In general, most of the clinicogenomic studies reduce the dimensionality of the data in a first step and then develop some predictive models on the selected features. A few studies merge these two steps into a single step taking the advantage of regularization based predictive models. Several statistical metrics were used to compare the performance of the combined clinicogenomic model with that of the clinical and genomic model. In most of the cases, the predictive power of the combined models was improved over that of individual clinical and genomic models, which justifies the usefulness of integration. However, in some cases, the combined model provided only marginal improvement; sometimes the performance of the genomic model was even worse than that of well-established clinical prognostic markers. This means that the value of traditional clinical variables should not be underestimated. Moreover, unlike genomic variables, the clinical variables are well established and validated through independent studies on multiple cohorts. Therefore, rigorous comparison between clinical and genomic variables is required in addition to looking at gains in predictive power. These observations motivated second sets of clinicogenomic studies which aimed at including the genomic variables into the prognosis models only if they provide some additional prognosis power. Thus, these models are biased towards clinical variables somehow. However, there are some additional issues with these kinds of approaches as well. If the models are biased too much towards the clinical variables, then the importance of genomic data may be subdued. This will hinder the discovery of potential new knowledge about complex diseases and thus may deviate from the main goal of elucidating new knowledge through integration. As a result, there is a trade-off between how much the combined model should be biased towards clinical dataset. Deciding this trade-off is not trivial. More systematic studies are required for this purpose.

Each of the data sources being integrated do not provide same amount of information, so the integration method should be cognizant of this difference in the datasets in terms of amount of information and the inherent properties in each datasets. Very few studies such as kernel based methods (Daemen, Gevaert et al. 2007) tried to preserve the individual properties available in each data source explicitly. However, this method used the vector based records only for both clinical and genomic data. On the other hand, the plethora of other types of medical, genetic and genomic data contains rich information with different types of structures such as time sequences, networks, replicates. Integrating such diverse type of data requires developing new computational techniques.

Interpretability of the obtained clinicogenomic models is a much desired property for personalized medicine as described in Section 0. However, predictive models mainly focus on improving the prediction power by combining the clinical and genomic data rather than interpretability. Therefore, most of the predictive models use those models that are more useful for improving the prediction power rather than producing interpretable models that can infer useful knowledge. Although tree based methods have been applied in this context, most of the studies applied more complex ensemble tree based models which are less interpretable than the original tree based rules. Moreover, separate dimensionality reduction step before developing clinico-genomic model may also reduce the interpretability of the model. For example, all these studies first combine the effect of genomic markers into a single score by either unsupervised techniques like PCA or separate pre-validation step, or some supervised techniques like PLS before developing any clinicogenomic model. The components do not provide information about the obtained genes and thus, the pathways involved in the disease progression, which is important for defining drug targets. Thus, incorporating the pre-validation framework for feature selection step is an open issue.

Disease heterogeneity is a very important challenge for clinico-genomic model development, but very few studies aim to handle all the challenges related to disease heterogeneity. For example, most studies try to find only different population groups using simple techniques, but do not form any detailed analysis regarding how those population groups exhibit different characteristics in terms of disease development. Moreover, further analysis is needed on whether there exist different subgroups of population, where each subgroup of population may have bias towards different datasets, i.e., each of these subgroups is associated with different types of markers. Such kinds of knowledge can elucidate new knowledge regarding how different types of markers can leads to different disease subtypes. Moreover, those subgroups of patients can be further verified for different clinical factors such as different demographic factors such as age, gender groups.

Most of the clinicogenomic models only aimed at utilizing two heterogeneous data sources during disease prognosis, but not elucidating the existing relationships between the clinical and genomic data. Different kinds of relationships between these two datasets can have different implication from domain perspectives. For example, if both datasets contain many correlated variables, then they contain similar types of information of a supplementary nature which cannot provide value to integrative studies. These types of correlations between datasets can be induced through other hidden factors [(Boulesteix, Porzelius et al. 2008)], e.g., the effect of a drug on gene expression [(Sotiriou and Piccart 2007)] during the treatment process. Besides correlation and independence, more complicated relationships like interactive or causal relationships may also exist between the clinical and genomic variables. For example, the intricate interaction between some genomic markers and environmental factors can make the disease phenotype more severe beyond their additive levels [(Loscalzo, Kohane et al. 2007), (Schadt 2009)], as mentioned in Section 1. Furthermore, there may be some genomic factors which have causal effects on some clinical variables. In that case, drugs can target those genomic variables in an early stage for better treatment design. For example, tumor surgery can be avoided if some causative genomic markers of tumor grade can be targeted in the early stage of breast cancer. Another interesting factor is that the clinical factors are not the causal factors of a disease phenotype unlike the genomic factors. Rather, most of the clinico-pathological variables are the observational properties of disease phenotype. Beside all these inter-dataset relationships, there may be also intra-dataset relationship among the variables within the same dataset representing the interactions or synergy between similar variables [(Braun, Cope et al. 2008),(Hwang, Sicotte et al. 2008)]. For example, familial hypertrophic cardiomyopathy is caused by mutations in several genes responsible for coding sarcomeric proteins [(Loscalzo, Kohane et al. 2007)], where each gene or protein is marginally inexplicable. Though some studies [(Boulesteix and Hothorn 2010)] seek to develop robust clinicogenomic model even in the presence of correlated variables, none of the clinicogenomic studies aim at elucidating this kind of inter- and intra- relationships between the clinical and genomic datasets. Further investigation is required to understand and utilize the potential broader relationships among different clinical and genomic variables when developing integrative models.

Very few studies validate the obtained clinicogenomic models extensively. Most of the studies did not compare the obtained model with other clinicogenomic models even those developed for the same disease. For example, only one out of seven breast cancer studies [(Sun, Goodison et al. 2007)] compared their final genomic signature with previous studies[(Van't Veer, Dai et al. 2002), (Wang, Klijn et al. 2005), (van de Vijver, He et al. 2002)]. Furthermore, very few of these studies were designed from methodological perspective. Most of the clinicogenomic studies applied different simple statistical and data-mining predictive models rather than applying and comparing different methodologies to get the best predictive model. Some studies [(Shedden, Taylor et al. 2008), (Bovelstad, Nygard et al. 2009)] tried to compare several dimensionality reduction techniques for regression based predictive model. On the other hand, the best way to integrate this uniform number of clinical and genomic variables after dimensionality reduction is not well understood. Some studies proposed intermediate integration for handling the challenges of heterogeneous data integration. Though in theory, kernel-SVM based intermediate integration is supposed to be more generalized, it did not provide significant improvement in the clinicogenomic context[(Daemen, Gevaert et al. 2007)]. Moreover, it is not clear how to represent the individual data using an intermediate format in the best possible way during intermediate integration. Kernel-based models also cannot find relationships among variables both within and across dataset. Alternatively, a graph-based approach can be utilized to address these issues. More systematic studies are required in this space to develop new methods to best leverage the diverse information available from both the clinical and gene expression data.

Although feature selection techniques have been explored quite a bit in the clinicogenomic context, there are still some issues when they are applied on gene expression datasets. First, the genomic features selected by dimensionality reduction techniques are not portable among different studies [(Chuang, Lee et al. 2007), (Naderi, Teschendorff et al. 2006)]. Second, often the topmost selected discriminating genes are hard to interpret and thus, do not provide any meaningful biological knowledge. Third and more importantly, in most of the cases, complex diseases are caused by either the interplay of a large group of oncogenes which have a combined effect on the overall disease or deregulation of a group of tumor-suppressor genes. For example, there are some well-known pathways that have been observed to be involved in disease progression such as the EGFR pathway, Wnt-signalling pathway, Hedgehog pathway, TGF-β, and so on. To address these issues, several methods for genome wide gene expression data [Gene set enrichment analysis [(Subramanian, Tamayo et al. 2005)], GSA [(Efron and Tibshirani 2007)], sub-GSE[(Yan and Sun 2008)]] have been designed to leverage prior genomic information. They look for pathways or groups of genes that have small individual discriminating power, but an overall large effect on disease progression. Similarly, epidemiologists already have some models in practice and they are only interested in incorporating new knowledge into their existing knowledge. Therefore, it would be advantageous to incorporate existing medical knowledge into the model development stage as prior knowledge so that it is easy for both validating and deploying the model. Incorporating such prior knowledge poses some challenges. For example, if some sort of dimensionality reduction is performed in the genomic data too rigorously upfront, then the small marginal affect can be ignored by the models which are hoping to elucidate information about pathways or synergistic relationships. On the other hand, if too many genomic variables are kept during the model developing, then the effect of low dimensional clinical variables can be underestimated [(Boulesteix and Sauerbrei 2011)]. The one-step approach can be an alternative solution, where no implicit dimensionality reduction is performed upfront and the model selects the optimum number of features. However, these kinds of sparse modeling techniques cannot take the prior domain knowledge into account easily. Developing new data mining algorithm to deal with these dimensionality reductions issues and incorporating prior knowledge needs further research.

Another important issue with most clinicogenomic studies is that most of these models consider only gene expression data as the genomic data from widely available public datasets. However, gene expression data contains information about transcriptional regulation only, and thus cannot provide any information about other aspects of complex cell mechanisms like post-transcriptional modification, protein synthesis and phosphorylation, copy number variation, random mutation in the genome and so on. Recent technological advancements have led to the advent of various high-throughput genomic data like protein abundance data, genome wide association (GWA) data, genetic interaction data, protein-protein interaction data, etc. It is important to note that these datasets are inherently related and each of them covers one particular aspect of cellular activity. Overlooking the inherent relationship could result in the discovery of biologically spurious associations, albeit statistically significant. For example, a gene that is differentially expressed can be spurious if the resultant protein is not differentially abundant due to post-transcriptional modifications. Integrating these enriched genomic data in the context of clinicogenomic studies pose further challenges. For example, the formats of other kind of genomic data are not uniform-gene expression or SNP are vector based, while PPI is graph based. Integrating these types of data with vector based clinical data is not trivial and needs further research.

Besides integrating clinical and genomic data sets, some multi-site clinicogenomic studies aim at integrating multiple similar types of datasets available from independent studies. However, these independent studies are performed in different experimental setups and different biological conditions, which might cause difference in probe design and final available gene expression profiles. Therefore, these issues have to be addressed with caution. In all multisite clinicogenomic studies, a simplistic approach was taken during integrating multiple genomic datasets by including only those genes as features that are significantly expressed (performed by t-test or other similar statistical test) in all cohorts. However, this reduces the number of features dramatically, because it is very less likely that genes will be simultaneously expressed in all independent cohorts. Moreover, it is biologically not meaningful, because different pathways may be disrupted for different groups of patients, even different groups of genes can be mutated for the same pathway during different environmental factors. So, it may be better to loosen the restriction a little bit to include genes that are not significantly expressed. More intelligently, genes can be selected from cohorts if they belong to a known pathway but do not meet the threshold of statistical significance. Moreover, more investigations are needed to handle the disparity in dimensions that arises when integrating them with low dimensional clinical variables.

Another strategy to reduce the experimental bias during the multisite studies can be subgroup analysis. If we assume that different subgroups of patients have different kinds of causal markers associated with them, then there may be heterogeneity even within a single cohort. This hypothesis of the existence of subgroups can be used to reduce the experimental bias in multi-site studies. Though there have been some efforts to develop subspace clinicogenomic models using association pattern mining algorithm, this needs further exploration in many dimensions.

In spite of great potential of clinico-genomic integration, the topic is still in a rudimentary phase. In general, integrating heterogeneous datasets like clinical and genomic data is a hard problem. The existing clinicogenomic models address these challenges partly. More detailed research is necessary especially to handle different kinds of relationships among variables and datasets; design robust model to handle disparate nature, structure, dimensionality, amount of information present in each dataset; incorporate prior knowledge into account; integrate diverse genomic and medical data besides gene expression and histo-pathological and demographic data; and finally validate the obtained clinicogenomic biomarkers rigorously in multiple independent cohort studies before final deployment for personalized medicine.

8 References

(03/16/2012). from www.adjuvantonline.com.

Acharya, C., D. Hsu, C. Anders, A. Anguiano, K. Salter, K. Walters, R. Redman, S. Tuchman, C. Moylan and S. Mukherjee (2008). "Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer." Jama **299**(13): 1574.

Agrawal, R., T. Imieli ski and A. Swami (1993). Mining association rules between sets of items in large databases, ACM.

Akaike, H. (1974). "A new look at the statistical model identification." IEEE transactions on automatic control **19**(6): 716-723.

Alizadeh, A., M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran and X. Yu (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature **403**(6769): 503-511.

Amaldi, E. and V. Kann (1998). "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems." Theoretical Computer Science **209**(1-2): 237-260.

Anastassiou, D. (2007). "Computational analysis of the synergy among multiple interacting genes." Molecular systems biology **3**(1).

Bair, E., T. Hastie, D. Paul and R. Tibshirani (2006). "Prediction by supervised principal components." Journal of the American Statistical Association **101**(473): 119-137.

Bay, S. D. and M. J. Pazzani (2001). "Detecting group differences: Mining contrast sets." Data Mining and Knowledge Discovery **5**(3): 213-246.

Beane, J., P. Sebastiani, T. Whitfield, K. Steiling, Y. Dumas, M. Lenburg and A. Spira (2008). "A prediction model for lung cancer diagnosis that integrates genomic and clinical features." Cancer Prevention Research **1**(1): 56.

Bebek, G., M. Koyutürk, N. D. Price and M. R. Chance (2012). "Network biology methods integrating biological data for translational science." Briefings in Bioinformatics.

Berlingerio, M., F. Bonchi, M. Curcio, F. Giannotti and F. Turini (2009). "Mining Clinical, Immunological, and Genetic Data of Solid Organ Transplantation." Biomedical Data and Applications: 211-236.

Bhattacharjee, A., W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno and M. Gillette (2001). "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." Proceedings of the National Academy of Sciences of the United States of America **98**(24): 13790.

Binder, H. and M. Schumacher (2008). "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models." BMC bioinformatics **9**(1): 14.

Bishop, C. M. and SpringerLink (2006). Pattern recognition and machine learning, Springer New York.

Biswal, B. B. and J. L. Ulmer (1999). "Blind source separation of multiple signal sources of fMRI data sets using independent component analysis." Journal of computer assisted tomography **23**(2): 265.

Blumberg, J. (2004). "PDA applications for physicians." ASCO News **16**: S4-S6.

Blute, M., E. Bergstralh, A. Iocca, B. Scherer and H. Zincke (2001). "Use of Gleason score, prostate specific antigen, seminal vesicle and margin status to predict biochemical failure after radical prostatectomy." The Journal of urology **165**(1): 119-125.

Boulesteix, A. and T. Hothorn (2010). "Testing the additional predictive value of high-dimensional molecular data." BMC bioinformatics **11**(1): 78.

Boulesteix, A., C. Porzelius and M. Daumer (2008). "Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value." Bioinformatics **24**(15): 1698.

Boulesteix, A. L. (2010). "Over-optimism in bioinformatics research." Bioinformatics **26**(3): 437.

Boulesteix, A. L. and W. Sauerbrei (2011). "Added predictive value of high-throughput molecular data to clinical data, and its validation." Breifings in bioinformatics.

Bovelstad, H., S. Nygard and Ø. Borgan (2009). "Survival prediction from clinico-genomic models-a comparative study." BMC bioinformatics **10**.

Braun, R., L. Cope and G. Parmigiani (2008). "Identifying differential correlation in gene/pathway combinations." BMC bioinformatics **9**(1): 488.

Breiman, L. (2001). "Random forests." Machine learning **45**(1): 5-32.

Buhlmann, P. and T. Hothorn (2007). "Boosting algorithms: Regularization, prediction and model fitting." Statistical Science **22**(4): 477-505.

Calhoun, V. D., T. Adali, G. Pearlson and K. Kiehl (2006). "Neuronal chronometry of target detection: fusion of hemodynamic and event-related potential data." Neuroimage **30**(2): 544-553.

Calnan, M. (2008). "Clinical uncertainty: is it a problem in the doctor-patient relationship?" Sociology of Health & Illness **6**(1): 74-85.

Campone, M., L. Campion, H. Roché, W. Gouraud, C. Charbonnel, F. Magrangeas, S. Minvielle, J. Genève, A. Martin and R. Bataille (2008). "Prediction of metastatic relapse in node-positive breast cancer: establishment of a clinicogenomic model after FEC100 adjuvant regimen." Breast cancer research and treatment **109**(3): 491-501.

Chi, J. T., Z. Wang, D. S. A. Nuyten, E. H. Rodriguez, M. E. Schaner, A. Salim, Y. Wang, G. B. Kristensen, Å. Helland and A. L. Børresen-Dale (2006). "Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers." PLoS medicine **3**(3): e47.

Chuang, H. Y., E. Lee, Y. T. Liu, D. Lee and T. Ideker (2007). "Network-based classification of breast cancer metastasis." Molecular systems biology **3**(1).

Cichocki, A., R. Zdunek, A. H. Phan and S. Amari (2009). Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation, Wiley.

Clarke, J. and M. West (2008). "Bayesian Weibull tree models for survival analysis of clinico-genomic data." Statistical methodology **5**(3): 238-262.

Cooper, G. and E. Herskovits (1992). "A Bayesian method for the induction of probabilistic networks from data." Machine learning **9**(4): 309-347.

Correa, N. M., T. Adali, Y. O. Li and V. D. Calhoun (2010). "Canonical correlation analysis for data fusion and group inferences." Signal Processing Magazine, IEEE **27**(4): 39-50.

Cover, T. and J. Thomas (2006). Elements of information theory, wiley.

Daemen, A., O. Gevaert and B. De Moor (2007). Integration of clinical and microarray data with kernel methods.

Dai, H., L. van't Veer, J. Lamb, Y. D. He, M. Mao, B. M. Fine, R. Bernards, M. van de Vijver, P. Deutsch and A. Sachs (2005). "A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients." Cancer research **65**(10): 4059-4066.

Dave, S., G. Wright, B. Tan, A. Rosenwald, R. Gascoyne, W. Chan, R. Fisher, R. Braziel, L. Rimsza and T. Grogan (2004). "Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells." New England Journal of Medicine **351**(21): 2159.

Delorme, A. and S. Makeig (2004). "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." Journal of neuroscience methods **134**(1): 9-21.

Dobbin, K. K., D. G. Beer, M. Meyerson, T. J. Yeatman, W. L. Gerald, J. W. Jacobson, B. Conley, K. H. Buetow, M. Heiskanen and R. M. Simon (2005). "Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays." Clinical cancer research **11**(2): 565.

Driouch, K., T. Landemaine, S. Sin, S. X. Wang and R. Lidereau (2007). "Gene arrays for diagnosis, prognosis and treatment of breast cancer metastasis." Clinical and Experimental Metastasis **24**(8): 575-585.

Duda, R. O., P. E. Hart and D. G. Stork (2001). Pattern classification, Citeseer.

Edén, P., C. Ritz, C. Rose, M. Fernö and C. Peterson (2004). "“Good Old” clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers." European journal of cancer **40**(12): 1837-1841.

Efron, B. and R. Tibshirani (2007). "On testing the significance of sets of genes." The Annals of Applied Statistics **1**(1): 107-129.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore and J. H. Nadeau (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nature Reviews Genetics **11**(6): 446-450.

Eifel, P., J. Axelson, J. Costa, J. Crowley, W. Curran Jr, A. Deshler, S. Fulton, C. Hendricks, M. Kemeny and A. Kornblith (2001). "National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000." Journal of the National Cancer Institute **93**(13): 979.

Ein-Dor, L., I. Kela, G. Getz, D. Givol and E. Domany (2005). "Outcome signature genes in breast cancer: is there a unique set?" Bioinformatics **21**(2): 171.

Ein-Dor, L., O. Zuk and E. Domany (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer, National Acad Sciences.

Fan, C., D. S. Oh, L. Wessels, B. Weigelt, D. S. A. Nuyten, A. B. Nobel, L. J. van't Veer and C. M. Perou (2006). "Concordance among gene-expression–based predictors for breast cancer." New England Journal of Medicine **355**(6): 560-569.

Fang, G., R. Kuang, G. Pandey, M. STEINBACH, C. MYERS and V. KUMAR (2010). "Subspace differential coexpression analysis: problem definition and a general approach." Pac Sympos Biocomput **15**: 145-156.

Fang, G., G. Pandey, W. Wang, M. Gupta, M. Steinbach and V. Kumar (2010). "Mining low-support discriminative patterns from dense and high-dimensional data." IEEE Transactions on Knowledge and Data Engineering.

Friedman, J., T. Hastie and R. Tibshirani (2000). "Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)." The annals of statistics **28**(2): 337-407.

Friedman, J. and B. Popescu (2004). Gradient directed regularization for linear regression and classification, Citeseer.

Futschik, M., M. Sullivan, A. Reeve and N. Kasabov (2003). "Prediction of clinical behaviour and treatment for cancers." Applied Bioinformatics **2**: 53-58.

Galea, M., R. Blamey, C. Elston and I. Ellis (1992). "The Nottingham Prognostic Index in primary breast cancer." Breast cancer research and treatment **22**(3): 207-219.

Gao, J., W. Fan, Y. Sun and J. Han (2009). Heterogeneous source consensus learning via decision propagation and negotiation, ACM.

Garber, M. E., O. G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. Van De Rijn, G. D. Rosen, C. M. Perou and R. I. Whyte (2001). "Diversity of gene expression in adenocarcinoma of the lung." Proceedings of the National Academy of Sciences of the United States of America **98**(24): 13784.

Gerds, T. and M. Schumacher (2006). "Consistent estimation of the expected Brier score in general survival models with right-censored event times." Biometrical Journal **48**(6): 1029-1040.

Gevaert, O., F. Smet, D. Timmerman, Y. Moreau and B. Moor (2006). "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks." Bioinformatics **22**(14): e184.

Goh, K., M. Cusick, D. Valle, B. Childs, M. Vidal and A. Barabási (2007). "The human disease network." Proceedings of the National Academy of Sciences **104**(21): 8685.

Goldhirsch, A., A. Coates, R. Gelber, J. Glick, B. Thürlimann and H. Senn (2006). "First—select the target: better choice of adjuvant treatments for breast cancer patients." Annals of Oncology **17**(12): 1772.

Goldhirsch, A., W. Wood, R. Gelber, A. Coates, B. Thurlimann and H. Senn (2003). "Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer." Journal of Clinical Oncology: 200304576.

Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing and M. Caligiuri (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science **286**(5439): 531.
Gönen, M. and E. Alpaydın (2011). "Multiple kernel learning algorithms." Journal of Machine Learning Research **12**: 2211-2268.
Graefen, M., P. Karakiewicz, I. Cagiannos, E. Klein, P. Kupelian, D. Quinn, S. Henshall, J. Grygiel, R. Sutherland and P. Stricker (2002). "Validation study of the accuracy of a postoperative nomogram for recurrence after radical prostatectomy for localized prostate cancer." Journal of Clinical Oncology **20**(4): 951.
Gui, J. and H. Li (2005). "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data." Bioinformatics **21**(13): 3001.
Guyon, I. and A. Elisseeff (2003). "An introduction to variable and feature selection." The Journal of Machine Learning Research **3**: 1157-1182.
Hamid, J. S., P. Hu, N. M. Roslin, V. Ling, C. M. T. Greenwood and J. Beyene (2009). "Data Integration in Genetics and Genomics: Methods and Challenges." Human Genomics.
Harrell Jr, F., R. Califf, D. Pryor, K. Lee and R. Rosati (1982). "Evaluating the yield of medical tests." Jama **247**(18): 2543.
Hastie, T., R. Tibshirani and J. H. Friedman (2009). The elements of statistical learning: data mining, inference, and prediction, Springer Verlag.
Heagerty, P., T. Lumley and M. Pepe (2000). "Time-dependent ROC curves for censored survival data and a diagnostic marker." Biometrics **56**(2): 337-344.
Herrero, J., R. Diaz-Uriarte and J. Dopazo (2003). "Gene expression data preprocessing." Bioinformatics **19**(5): 655-656.
Hoerl, A. E. and R. W. Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems." Technometrics **12**(1): 55-67.
Hoeting, J., D. Madigan, A. Raftery and C. Volinsky (1999). "Bayesian model averaging: A tutorial." Statistical Science **14**(4): 382-401.
Höfling, H. and R. Tibshirani (2008). "A study of pre-validation." Annals **2**(2): 643-664.
Hothorn, T., P. Buhlmann, S. Dudoit, A. Molinaro and M. Van Der Laan (2005). "Survival ensembles." Biostatistics.
Hristidis, V. (2009). Information Discovery on Electronic Health Records, Chapman & Hall.
Huang, E., S. H. Cheng, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, A. Bild, E. S. Iversen, M. Liao and C. M. Chen (2003). "Gene expression predictors of breast cancer outcomes." The Lancet **361**(9369): 1590-1596.
Huang, E., S. Ishida, J. Pittman, H. Dressman, A. Bild, M. Kloos, M. D'Amico, R. G. Pestell, M. West and J. R. Nevins (2003). "Gene expression phenotypic models that predict the activity of oncogenic pathways." Nature genetics **34**(2): 226-230.
Hwang, T. H., H. Sicotte, Z. Tian, B. Wu, J. P. Kocher, D. A. Wigle, V. Kumar and R. Kuang (2008). "Robust and efficient identification of biomarkers by classifying features on graphs." Bioinformatics **24**(18): 2023.
Ioannidis, J. (2005). "Microarrays and molecular research: noise discovery?" Lancet **365**(9458): 454.
Jolliffe, I. (2002). "Principal component analysis."
Kammers, K., M. Lang, J. G. Hengstler, M. Schmidt and J. Rahnenführer (2011). "Survival models with preclustered gene groups as covariates." BMC bioinformatics **12**(1): 478.
Kaplan, E. and P. Meier (1958). "Nonparametric estimation from incomplete observations." Journal of the American Statistical Association **53**(282): 457-481.
Kasabov, N. (2001). "On-line learning, reasoning, rule extraction and aggregation in locally optimized evolving fuzzy neural networks." Neurocomputing **41**(1-4): 25-45.

Kato, G. J., M. T. Gladwin and M. H. Steinberg (2007). "Deconstructing sickle cell disease: reappraisal of the role of hemolysis in the development of clinical subphenotypes." Blood reviews **21**(1): 37-47.

Kattan, M., T. Wheeler and P. Scardino (1999). "Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer." Journal of Clinical Oncology **17**(5): 1499.

Kelley, R. and T. Ideker (2005). "Systematic interpretation of genetic interactions using protein networks." Nature biotechnology **23**(5): 561-566.

Kittler, J. (1998). "Combining classifiers: A theoretical framework." Pattern Analysis & Applications **1**(1): 18-27.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, Citeseer.

Kohavi, R. and G. John (1997). "Wrappers for feature subset selection." Artificial intelligence **97**(1-2): 273-324.

Konstantinopoulos, P. A., S. A. Cannistra, H. Fountzilas, A. Culhane, K. Pillay, B. Rueda, D. Cramer, M. Seiden, M. Birrer and G. Coukos (2011). "Integrated analysis of multiple microarray datasets identifies a reproducible survival predictor in ovarian cancer." PloS one **6**(3): e18202.

Kumar, U., S. Grigorakis, H. Watt, R. Sasi, L. Snell, P. Watson and S. Chaudhari (2005). "Somatostatin receptors in primary human breast cancer: quantitative analysis of mRNA for subtypes 1–5 and correlation with receptor protein expression and tumor pathology." Breast cancer research and treatment **92**(2): 175-186.

Lê Cao, K. A., P. G. P. Martin, C. Robert-Granié and P. Besse (2009). "Sparse canonical methods for biological data integration: application to a cross-platform study." BMC bioinformatics **10**(1): 34.

Lê Cao, K. A., E. Meugnier and G. J. McLachlan (2010). "Integrative mixture of experts to combine clinical factors and gene markers." Bioinformatics **26**(9): 1192-1198.

Li, J. and Y. Sun (2006). Iterative relief for feature weighting: algorithms, theories and applications. ICML.

Li, K. (1991). "Sliced inverse regression for dimension reduction." Journal of the American Statistical Association **86**(414): 316-327.

Li, L. (2006). "Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information." Bioinformatics **22**(4): 466.

Li, L., L. Chen, D. Goldgof, F. George, Z. Chen, A. Rao, J. Cragun, R. Sutphen and J. Lancaster (2005). "Integration of clinical information and gene expression profiles for prediction of chemo-response for ovarian cancer."

Liew, A. W. C., N. F. Law and H. Yan (2010). "Missing value imputation for gene expression data: computational techniques to recover missing data from available information." Briefings in Bioinformatics.

Liu, R., X. Wang, G. Y. Chen, P. Dalerba, A. Gurney, T. Hoey, G. Sherlock, J. Lewicki, K. Shedden and M. F. Clarke (2007). "The prognostic role of a gene signature from tumorigenic breast-cancer cells." New England Journal of Medicine **356**(3): 217-226.

Loscalzo, J., I. Kohane and A. L. Barabasi (2007). "Human disease classification in the postgenomic era: a complex systems approach to human pathobiology." Molecular systems biology **3**(1).

Ma, S. and J. Huang (2007). "Combining Clinical and Genomic Covariates via Cov-TGDR." Cancer Informatics **3**: 371.

Maojo, V. and F. Martin-Sanchez (2004). "Bioinformatics: towards new directions for public health." METHODS OF INFORMATION IN MEDICINE. **43**(3): 208-214.

McClellan, J. and M. C. King (2010). "Genetic heterogeneity in human disease." Cell **141**(2): 210-217.

Meingast, M., T. Roosta and S. Sastry (2008). Security and privacy issues with health care information technology, IEEE.

Naderi, A., A. Teschendorff, N. Barbosa-Morais, S. Pinder, A. Green, D. Powe, J. Robertson, S. Aparicio, I. Ellis and J. Brenton (2006). "A gene-expression signature to predict survival in breast cancer across independent data sets." Oncogene **26**(10): 1507-1516.

Nevins, J., E. Huang, H. Dressman, J. Pittman, A. Huang and M. West (2003). "Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction." Human molecular genetics **12**(Review Issue 2): R153.

Obulkasim, A., G. A. Meijer and M. A. van de Wiel (2011). "Stepwise classification of cancer samples using clinical and molecular data." BMC bioinformatics **12**(1): 422.

Oelker, M. R. and A. L. Boulesteix "On the Simultaneous Analysis of Clinical and Omics Data-a Comparison of Globalboosttest and Pre-validation Techniques."

Oliver, J. and D. Hand (1995). On pruning and averaging decision trees, Citeseer.

Paoli, S., G. Jurman, D. Albanese, S. Merler and C. Furlanello (2008). "Integrating gene expression profiling and clinical data." International Journal of Approximate Reasoning **47**(1): 58-69.

Park, M. Y. and T. Hastie (2007). "L1-regularization path algorithm for generalized linear models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69**(4): 659-677.

Partin, A., J. Mohler, S. Piantadosi, C. Brendler, M. Sanda, P. Walsh and J. Epstein (1995). "Selection of men at high risk for disease recurrence for experimental adjuvant therapy following radical prostatectomy*." Urology **45**(5): 831-838.

Pavlidis, P., J. Weston, J. Cai and W. Grundy (2001). Gene functional classification from heterogeneous data, ACM.

Pittman, J., E. Huang, H. Dressman, C. Horng, S. Cheng, M. Tsou, C. Chen, A. Bild, E. Iversen and A. Huang (2004). "Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes." Proceedings of the National Academy of Sciences of the United States of America **101**(22): 8431.

Potti, A., S. Mukherjee, R. Petersen, H. K. Dressman, A. Bild, J. Koontz, R. Kratzke, M. A. Watson, M. Kelley and G. S. Ginsburg (2006). "A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer." New England Journal of Medicine **355**(6): 570.

Raftery, A., D. Madigan and J. Hoeting (1997). "Bayesian model averaging for linear regression models." Journal of the American Statistical Association **92**(437): 179-191.

Redmond Jr, D. E. (1970). "Tobacco and cancer: the first clinical report, 1761." New England Journal of Medicine **282**(1): 18-23.

Royston, P. and W. Sauerbrei (2004). "A new measure of prognostic separation in survival data." Statistics in medicine **23**(5): 723-748.

Schadt, E. E. (2009). "Molecular networks as sensors and drivers of common human diseases." Nature **461**(7261): 218-223.

Schumacher, M., H. Binder and T. Gerds (2007). "Assessment of survival prediction models based on microarray data." Bioinformatics **23**(14): 1768.

Schwarz, E., F. Leweke, S. Bahn and P. Liò (2009). "Clinical bioinformatics for complex disorders: a schizophrenia case study." BMC bioinformatics **10**(Suppl 12): S6.

Shedden, K., J. Taylor, S. Enkemann, M. Tsao, T. Yeatman, W. Gerald, S. Eschrich, I. Jurisica, T. Giordano and D. Misek (2008). "Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study." Nature medicine **14**(8): 822-827.

Shipp, M., D. Harrington, J. Anderson, J. Armitage, G. Bonadonna, G. Brittinger, F. Cabanillas, G. Canellos, B. Coiffier and J. Connors (1993). "A predictive model for aggressive non-Hodgkin's lymphoma. The International non-Hodgkin's Lymphoma Prognostic Factors Project." N Engl J Med **329**(14): 987-994.

Šilhavá, J. and P. Smrz (2009). "Additional predictive value of microarray data compared to clinical variables." 4th IAPR International Conference on Pattern Recognition in Bioinformatics

Simon, R., M. D. Radmacher, K. Dobbin and L. M. McShane (2003). "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification." Journal of the National Cancer Institute **95**(1): 14.
Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico and J. P. Richie (2002). "Gene expression correlates of clinical prostate cancer behavior." Cancer cell **1**(2): 203-209.
Smialowski, P., D. Frishman and S. Kramer (2010). "Pitfalls of supervised feature selection." Bioinformatics **26**(3): 440.
Sotiriou, C. and M. Piccart (2007). "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?" Nature Reviews Cancer **7**(7): 545-553.
Sotiriou, C., P. Wirapati, S. Loi, B. Haibe-Kains, C. Desmedt, A. Tutt, P. Ellis, M. Buyse, M. Delorenzi and M. Piccart (2006). "Comprehensive analysis integrating both clinicopathological and gene expression data in more than 1500 samples: Proliferation captured by gene expression grade index appears to be the strongest prognostic factor in breast cancer (BC)." J Clin Oncol **24**: 4S.
Sotiriou, C., P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz and B. Haibe-Kains (2006). "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis." JNCI Cancer Spectrum **98**(4): 262.
Spira, A., J. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y. Dumas, P. Calner and P. Sebastiani (2007). "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer." Nature medicine **13**(3): 361-366.
Stephenson, A., A. Smith, M. Kattan, J. Satagopan, V. Reuter, P. Scardino and W. Gerald (2005). "Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy." Cancer **104**(2): 290.
Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub and E. S. Lander (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences of the United States of America **102**(43): 15545.
Sui, J., T. Adali, G. Pearlson, H. Yang, S. R. Sponheim, T. White and V. D. Calhoun (2010). "A CCA+ ICA based model for multi-task brain imaging data fusion and its application to schizophrenia." Neuroimage **51**(1): 123-134.
Sun, Y., S. Goodison, J. Li, L. Liu and W. Farmerie (2007). "Improved breast cancer prognosis through the combination of clinical and genetic markers." Bioinformatics **23**(1): 30.
Szklo, M. (1998). "Population-based cohort studies." Epidemiologic reviews **20**(1): 81.
Tan, P., M. Steinbach and V. Kumar (2006). Introduction to data mining, Pearson Addison Wesley Boston.
Teschendorff, A., A. Naderi, N. Barbosa-Morais, S. Pinder, I. Ellis, S. Aparicio, J. Brenton and C. Caldas (2006). "A consensus prognostic gene expression classifier for ER positive breast cancer." Genome Biol **7**(10): R101.
Teschendorff, A. E., A. Miremadi, S. E. Pinder, I. O. Ellis and C. Caldas (2007). "An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer." Genome Biol **8**(8): R157.
Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological): 267-288.
Tibshirani, R., B. Efron and S. U. D. o. Biostatistics (2002). Pre-validation and inference in microarrays, Stanford University, Department of Biostatistics.
Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**(6): 520.

Troyanskaya, O. G., K. Dolinski, A. B. Owen, R. B. Altman and D. Botstein (2003). "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)." Proceedings of the National Academy of Sciences **100**(14): 8348.
Truntzer, C., D. Maucort-Boulch and P. Roy (2008). "Comparative optimism in models involving both classical clinical and gene expression information." BMC bioinformatics **9**(1): 434.
Tsiliki, G. and S. Kossida (2011). "Fusion methodologies for biomedical data." Journal of Proteomics.
Tutz, G. and H. Binder (2007). "Boosting ridge regression." Computational Statistics & Data Analysis **51**(12): 6044-6059.
Ulitsky, I., R. Karp and R. Shamir (2008). Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles, Springer.
Van't Veer, L., H. Dai, M. Van de Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven and C. Roberts (2002). "Gene expression profiling predicts clinical outcome of breast cancer."
van de Vijver, M., Y. He, L. van't Veer, H. Dai, A. Hart, D. Voskuil, G. Schreiber, J. Peterse, C. Roberts and M. Marton (2002). "A gene-expression signature as a predictor of survival in breast cancer." The New England journal of medicine **347**(25): 1999.
Van Dongen, S. (2000). "A cluster algorithm for graphs." Report-Information systems(10): 1-40.
van Vliet, M. H., H. M. Horlings, M. J. van de Vijver, M. J. T. Reinders and L. F. A. Wessels (2012). "Integration of Clinical and Gene Expression Data Has a Synergetic Effect on Predicting Breast Cancer Outcome." PloS one **7**(7): e40358.
Vapnik, V. N. (2000). The nature of statistical learning theory, Springer Verlag.
Wang, S., L. Ooi and K. Hui (2007). "Identification and validation of a novel gene signature associated with the recurrence of human hepatocellular carcinoma." Clinical cancer research **13**(21): 6275.
Wang, Y., J. Klijn, Y. Zhang, A. Sieuwerts, M. Look, F. Yang, D. Talantov, M. Timmermans, M. Meijer-van Gelder and J. Yu (2005). "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." The Lancet **365**(9460): 671-679.
Weinberg, R. A. (2007). The biology of cancer, Garland Science.
West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks and J. Nevins (2001). "Predicting the clinical status of human breast cancer by using gene expression profiles." Proceedings of the National Academy of Sciences of the United States of America **98**(20): 11462.
West, M., G. Ginsburg, A. Huang and J. Nevins (2006). "Embracing the complexity of genomic data for personalized medicine." Genome research **16**(5): 559.
Westra, B. L., S. Dey, G. Fang, M. Steinbach, V. Kumar, C. Oancea, K. Savik and M. Dierich (2011). "Interpretable Predictive Models for Knowledge Discovery from Home-Care Electronic Health Records." Journal of Healthcare Engineering **2**(1): 55-74.
Wold, H. (1985). "Partial least squares."
Xu, L., G. Pearlson and V. D. Calhoun (2009). "Joint source based morphometry identifies linked gray and white matter group differences." Neuroimage **44**(3): 777-789.
Yan, X. and F. Sun (2008). "Testing gene set enrichment for subset of genes: Sub-GSE." BMC bioinformatics **9**(1): 362.
Zhou, W., G. Liu, D. P. Miller, S. W. Thurston, L. L. Xu, J. C. Wain, T. J. Lynch, L. Su and D. C. Christiani (2002). "Gene-environment interaction for the ERCC2 polymorphisms and cumulative cigarette smoking exposure in lung cancer." Cancer research **62**(5): 1377-1381.

Appendix:

A. *Dimensionality reduction technique for clinicogenomic studies:* We found both feature selection and feature extraction techniques from machine learning have been applied for dimensionality reduction in clinicogenomic model development. We will discuss both of these kinds of approach as well as some more other techniques like clustering and pre-selecting genes based on incorporating prior knowledge. The detailed application of such techniques is described in the next section.

1. Feature extraction techniques: Feature extraction methods tries to extract a reduced set of features to decrease the redundancy of the data. Some integration studies first developed their own dimensionality reduction technique modifying the existing technique and then, applied their own model in the context of clinicogenomic model. So, the main perspective of these studies was to design machine-learning based dimensionality reduction techniques for clinicogenomic data. For example, Li et al. [(Li 2006)] developed a dimensionality reduction technique to select a reduced sized-set of genes consisting of linear combination of genes by applying two step dimensionality reduction techniques: the unsupervised principal component analysis (PCA) method followed by a supervised dimensionality reduction method called sliced inverse regression (SIR)[(Li 1991)] which searches for the smallest subspace of the dataset preserving the regression information of outcome variable(**y**) given the input feature space(**X**). Finally, they modified the SIR method to partial SIR (PSIR) so that it can include clinical variables as the mandatory variables without applying dimensionality reduction on it. The main advantage of this model is that it does not assume any kind of prior distribution on the model. However, it requires linearity assumption on the marginal distribution of **X.** This method is also difficult to implement. Moreover, the main problem of this method is that it falls into the category of feature extraction where the final predictive model is built on the extracted features which are the linear combination of features. So, it is hard for clinicians to interpret those extracted features so that they can be used for further use like drug targets.

2. Feature selection methods: To overcome the limitations of such feature extraction methods, feature selection methods seem the most popular technique in this context of building clinicogenomic model. Feature selection methods try to select a subset of existing features based on their predictability so that most irrelevant and redundant genes are removed. This technique is different than feature extraction which aims at extracting new features from the existing new variable, and thus feature selection can provide a useful explanation about the selected features describing which variables are the most important one and how they are related to each other. There are three supervised strategies for feature selections: filtering method, wrapper method and embedded method. In next three paragraphs, we will provide the discussion and some examples of these three methods in biological domain.

a) Filtering methods: Filtering criteria ranks all the features based on some metric and then, selects those features that achieve those score. Among several available metric for evaluating such features, statistical tests like student t-test, Wilcoxon non-parametric rank sum test, Fishar's exact test seem mostly used in the medical domain. The most popular statistics used for pre-selecting the genes for developing clinicogenomic model is the t-test. For example, Lihua Li et. al.[(Li, Chen et al. 2005)], Stephenson et. al. [(Stephenson, Smith et al. 2005)], Silhava et al.[(Šilhavá and Smrz 2009)] performed two-sample t-test between the disease recurrent and non-recurrent tissue sample. Though t-test is the most popular filtering method, it assumes that the data is normally distributed. This assumption may not be valid for all genomic dataset. Several alternative non-parametric tests exist which do not make any such assumptions of the underlying data. Daemen et al.[(Daemen, Gevaert et al. 2007)] used such a distribution-free test statistics called Wilcoxon rank sum test to pre-select top 1000 most significant genes to build a kernel based integrated clinicogenomic model. Some other study like Wang et. al [(Wang, Ooi et al. 2007)] performed both parametric and non-parametric tests for pre-selecting the genes. In some another recent studies, for example integrative study by Campone et al.[(Campone, Campion et al. 2008)], univariate cox regression itself was used to select the genes using hazard ratio as the filtering criteria. In addition, Campone et al. 2008 selected those genes which were selected most commonly by univariate cox regression in several random samplings of the data to remove the selection bias. However, sometimes even after the correction for multiple hypothesis testing, these test statistics provide lot of statistically significant genes which may not be interesting to investigate further from domain perspectives. Therefore, besides using simply tests, some studies also selected genes using some intestingness measures from domain knowledge like the level of fold change, the number of samples where the gene is expressed significantly. For example, Wang et al.[(Wang, Ooi et al. 2007)] and Stephenson et al. [(Stephenson, Smith et al. 2005)] selected those genes which have significant fold change >1.5. In some other studies [Mario et al. 2008, Campone et al. 08], principal component analysis was performed in a follow-up step to provide further reductions in dimensionality and thus all genes selected by univariate Cox regression are represented by topmost principal component. However, this also shares the same kind of disadvantages described earlier for feature extraction.

These statistical test based filtering methods have several advantages like they are easier to implement, faster than other sophisticated feature selection methodology, and easier to interpret and verify for clinicians. Despite these advantages, they suffer by some limitations in the context of gene selection [(Sun, Goodison et al. 2007)]. First, filter methods miss the interaction among the features, as it evaluates each gene independently. Some features may have small individual marginal effects with the disease outcome variable, while they might have greater effects when they are considered in a combination. This is a big concern especially for genomic datasets, because genes are normally expressed by different biological processes and pathways which also regulate other genes at a time. Second, Filter methods are unable to remove some of the redundant features sometime. For example, filter methods cannot assign separate scores for co-regulated genes. Such irrelevant features are undesirable from both computational and clinical perspectives. From computational perspective, such irrelevant features may deteriorate the performance of the classifiers [(Kohavi and John 1997)]. Moreover from clinical perspective, the examination of the expression levels of redundant genes may increase the clinical cost. To overcome these disadvantages of filter methods different alternative computational methods have been proposed in the community of machine learning and data mining for feature selection. Some of those methods have been also incorporated into the integrative study of clinical and genomic data as well as described in the next two sections.

b) Wrapper method: Wrapper models search for an optimal subset of features in the whole feature space and then evaluate the selected subset based on some model built on it. However, searching for such an optimal subset of features poses challenges for learning algorithms specially, exhaustive search for substantial features is quite impossible for high-dimensional genomic data. This is because the possible search space increases exponentially with the increase of dimensionality. That's why most existing solutions depend on heuristic combinatorial search with a possible trade-off for optimality. Most of such wrapper methods rely on some greedy search techniques and stop the search based on some objective functions or optimality criteria [either wiki or something else]. Moreover, it is also difficult sometimes to define an appropriate objective functions which can be optimized easily for designing faster search criterion [(Sun, Goodison et al. 2007)].

Sun et. al. (Li and Sun 2006) recently proposed a new objective function in their proposed feature selection techniques called I-RELEIF to overcome the limitations of both filters methods and the combinatorial search problem of wrapper methods. It assigns weights to the features based on the block distance of each sample to its nearest neighbor and then tries to optimize those weights using analytical and numerical solutions such that the accuracy of the nearest-neighbor classifier of each sample in the original feature space is maximized. The detailed description about how it was applied in the context of clinicogenomic model development is given in the next section.

Both of the filtering method and wrapper method fall in the general category of supervised method as they consider the class labels into account while selecting the features. Furthermore in a following step, those selected genes were fed into a classifier model along with clinical variable to build a combined clinicogenomic model along with cross-validation for estimating the performance of the models. Such kind of two step approach will produce biased classification results unless this dimensionality reduction is performed for each training data within each iteration of the cross-validation. Very few studies [(Sun, Goodison et al. 2007), (Stephenson, Smith et al. 2005)] performed the gene selection within the cross-validation loop with LOOCV and so, was able to produce unbiased classification results. Such kind of mistake while feature selection and classification are performed together was reported also in other genomic studies (Smialowski, Frishman et al. 2010)]. Moreover, all of these dimensionality reduction techniques try to reduce dimensionality only on gene expression data independent of clinical variables So, these methods did not consider reducing the dimensionality of the clinical dataset at the same time rather most of the study relied on some previously developed clinical index like IPI (International Prognostic index developed for Large B-Cell Lymphoma[(Shipp, Harrington et al. 1993)]) instead of original clinical database. Regularization based classifiers can be an alternative solution for these potential problems and they are described in next section in the context of integrative clinicogenomic model development.

c) Penalized statistical learning methods: Recently, several techniques have been developed in machine learning like boosting or regularization-based penalized algorithms where feature selection is performed simultaneously with the developing of predictive model. Such kind of embedded system can be a great solution to the above mentioned two step approaches, as it will perform both feature selection and outcome prediction at the same time. The penalized methods introduce a penalty term for providing shrinkage in the model development which is learnt from the data as well. Some of the initial studies of such penalized model building for genomic data were investigated by Gui and Li et al [(Gui and Li 2005)] by developing L-1 penalized Cox regression model for gene expression data [provide more citations]. However, most of the currently available penalized methods that have been applied into genomic data treat all of the covariates with same weights. In the context of integrated clinicogenomic model development, putting same weights for all the variables may not be a good idea. Because the inherent nature of the datasets is totally different, for example, one unit change in gene expression has different interpretation than one unit change in clinical variable [Ma et al.]. Moreover, it is more desirable to put more penalties on gene expression data than clinical variables because, gene expression data contains more irrelevant features than clinical variables and the dimensionality is also higher in genomic data in compare to clinical variables. Moreover, most of the clinical variables are selected from domain knowledge like pathological and demographic variables which are already being used for clinical predictions. So, dimensionality reduction on those variables may not be desirable as well.

For these reasons, some of the integration efforts incorporated the regularization methods into integration study with a possible modification of assigning different weights to clinical and genomic data, so that those two types of data can be regularized differently. For example, Binder et al. 2008 incorporated their own technique developed previously (this->3) in this context of developing combined clinicogenomic model with a modification. They developed a gradient boosting method for Cox proportional hazard model in order to allow flexible penalty structure so that dimensionality reduction is performed in the genomic data only and the mandatory clinical variables are included without any penalization. With similar motivation, Ma et al. 2007 extended the original gradient search based Threshold Gradient Directed Regularization (TGDR) approach [Friedman and Popescu 2004] into their Covariate-adjusted TGDR (Cov-TGDR), where two more parameters were introduced to represent the regularization co-efficient for the two corresponding datasets of clinical and genomic data. This provides a more generalized framework in compare to the former approach as dimensionality reduction can be performed in both clinical and genomic datasets.

Beside all these statistical methods of feature selection and extraction described so far, some studies incorporated the concept of metagenes which are defined by the clusters of similar genes. Such types of metagenes were developed using *clustering* approach from data mining community in previous studies. In the next section we will describe some of these studies where clustering based approach has been applied to select some of the genes.

3. Clustering based approach: Metagenes are defined as the representative signature of similar genes [(Wang, Klijn et al. 2005)]. Those metagenes are obtained by performing clustering to gene expression profile for finding groups of genes with coherent expressions over all of the samples. Nevins et al. 2003 developed one of the earliest clinicogenomic models using such clustering based metagene approach. In this study, they applied K-mean clustering first to get some clusters of genes with similar profile and then they computed the weighted average of the genes within each clusters to join them into a single meta-genes. In an extended study of this initial study, Pittman et al. 2004 computed the metagenes as the first principal components of those clusters of genes created by K-means clustering. However, as the true number of the clusters is unknown here, it was hard to guess the correct number of clusters upfront which is required for K-mean algorithm. This motivated Clarke et al. 2008 to generate large number of clusters first and then filter out those clusters that are not true representative of metagenes in a post processing steps. For pruning such spurious clusters, they performed some sort of post-validation by silhouette widths for the genes within the same cluster in order to get the statistical significance which is determined by a permutation-based null distribution. Besides K-Mean clustering other types of clustering like hierarchical clustering, DBSCAN and so on [(Tan, Steinbach et al. 2006)], which does not require the number of clusters to be specified as apriori, can be investigated also for this purpose. One big advantage of clustering based methods is that each cluster can capture the expressions of several genes belonging to that cluster and so can retain more information during classification as confirmed by the superior results of the clinico-genomic study by Shedden et al.[(Shedden, Taylor et al. 2008)].

4. Pre-selecting genes from previous knowledge: Some of the studies relied on the previous study or prior biological knowledge to select the most appropriate genes to determine the most relevant genes for the purpose of the study, rather than only depending on the dimensionality reduction techniques described above. For example, Beane et al. 2008 used a gene expression index developed using majority voting algorithm [(Golub, Slonim et al. 1999)] in a previous study. Similarly, Teschendorff et al. 2006 performed an extended study on six breast cancer datasets including his own collected samples but relied only on the genes that are common between their own dataset with all of the previous five studies. In a similar study as Teschendorff et al. 2006, Acharya et al. 2009 also tried to discover the underlying biological behaviors of different disease subgroups of breast cancer. In this effort of combining five different breast cancer datasets, they only selected those gene signatures of altered tumor microenvironment states, along with oncogenic pathway deregulation, and chemotherapy response. More specifically, genes involved in chromosomal instability, wound healing, IGS, epigenetic stem cells, and tumor necrosis factor α (TNF-α) which are believed to have some effects in breast cancer survival were included for further study. Similarly, during the finding risk factors of common liver diseases, Berlingerio et al [(Berlingerio, Bonchi et al. 2009)] also selected gene expressions of six sites from HLA which are already known to have effects on the liver diseases. Though such kind of biological knowledge from previous studies provide a way to increase the confidence and relevance of selecting genes, this narrows down the possibility of selecting new genes that are not described in previous literature.

All of these above mentioned feature selection techniques have their own advantages and disadvantages. So, some studies examined several feature selection techniques rather than confined in one feature selection method, and then use that feature selection method which produces the best results for the classification methods in the follow-up studies. For example, Shedden et al. 2008[(Shedden, Taylor et al. 2008)] used several methods like gene clustering, filtering based on univariate regression models, etc. for selecting genes and got the best classification results for clustering methods. In another recent study, Bovelstad et al. 2009 built compared seven different approaches for dimensionality reduction covering several techniques already described in last few paragraphs for building Cox regression model. The details of all these approaches are described more elaborately in the next section where Cox based models are discussed.

B: Preprocessing of data

Most clinicogenomic studies perform several preprocessing steps before developing clinicogenomic integrative models (Herrero, Diaz-Uriarte et al. 2003). The goal of the preprocessing of the datasets is to reduce the heterogeneity between the two datasets as much as possible, so that model development on the combined dataset becomes as easy as possible. For example, genomic data contains more noise and missing values than clinical variables. Many clinicogenomic studies incorporate the same preprocessing techniques as many genomic studies. For example, filtering based on the fold change to drop some irrelevant genes having very few expression labels throughout the samples was used by (Shedden, Taylor et al. 2008). Another common preprocessing step is to normalize the data to reduce the sample variances (Shedden, Taylor et al. 2008). Many of the microarray studies contain missing values for certain probes. Several studies (Li 2006), (Binder and Schumacher 2008), (Clarke and West 2008) also addressed this issue by using imputation techniques e.g., the nearest neighbor algorithm (Troyanskaya, Cantor et al. 2001), (Liew, Law et al. 2010) before developing any clinicogenomic model. Another challenge for binary classification on any medical domain is the imbalanced class problem, as disease patients are rare than healthy controls. Special corrections are needed for handling imbalanced classes to avoid the bias of the model towards the majority class. Very few clinicogenomic studies (Stephenson, Smith et al. 2005) are aware of this fact.

The biggest challenge that these heterogeneous datasets pose is the disparate dimensionalities of the clinical and genomic data. Since clinical data is low dimensional, most of the two-step approaches first reduces the dimensionality of the high-dimensional genomic data before building any predictive model. Dimensionality reduction (DR) [(Guyon and Elisseeff 2003), (Xing, Jordan et al. 2001), (Bernau and Boulesteix 2010)] is a widely studied research topic in the area of from machine learning. There are various advantages of dimensionality reduction: reducing the overfitting problem and thus improving the prediction power of the models, reducing the time and space complexity of model development, enhancing the understanding of generative factors of the data, and facilitating visualization (Guyon and Elisseeff 2003). In general, DR techniques can be divided into two main categories: *feature selection* and *feature extraction*. Feature selection techniques select a subset of important variables retaining as much information as possible from the data, while feature extraction techniques create some new features from the existing variables. Another categorization of the dimensionality reduction techniques can be based on whether the label information is used. In this section, we will provide a brief summarization of all techniques that have been used by clinicogenomic models. A detailed description of all these techniques has been included in the Appendix A.

Most of the feature extraction techniques such as principal component analysis (PCA) (Jolliffe 2002) creates new features which are a linear transformation of the data for example. Creating features from some natural groups in the data through clustering is another technique (Pittman, Huang et al. 2004, Clarke and West 2008). For example, each cluster centroid or first principal component of each cluster can be represented as a new feature. All these approaches do not use the label information during constructing features. Several clinicogenomic studies also aimed to modify the original unsupervised feature creation techniques. For example, Boulesteix et al. (Boulesteix, Porzelius et al. 2008) used supervised PLS techniques to defines genomic features like PCA. On the other hand, Lexin Li [] used a two-step dimension reduction technique: unsupervised principal component analysis (PCA) method followed by a supervised dimensionality reduction method called sliced inverse regression (SIR) (Li 1991). One big advantage of feature creation as opposed to feature selection is that each extracted feature can retain more information about the data. For example, the metagenes (West, Blanchette et al. 2001), (Huang, Ishida et al. 2003), (Huang, Cheng et al. 2003) that are co-expressed in a relevant pathway can be summarized by a feature represented by a single cluster. However, one major problem with these feature creation techniques is that the features are not biologically interpretable, which is utmost important in the context of biomarker discovery.

In the domain of building a clinicogenomic integrative model, feature selection seems the most popular technique, because of interpretability. In general, three types of feature selection techniques are used: filtering method, wrapper method and embedded methods. Note that all these methods are supervised in the nature. Filtering methods rank all the features based on various metrics and then, selects those features that achieve a particular score. Some of the metrics used in the clinicogenomic context are parametric (e.g., student t-test (Li, Chen et al. 2005), (Stephenson, Smith et al. 2005), (Šilhavá and Smrz 2009)), non-parametric statistical tests (e.g., Wilcoxon rank sum test (Daemen, Gevaert et al. 2007),(Wang, Ooi et al. 2007)), univariate Cox regression(Campone, Campion et al. 2008) and domain knowledge like the level of fold change (Wang, Ooi et al. 2007), (Stephenson, Smith et al. 2005), (Höfling and Tibshirani 2008), the number of samples where the gene is expressed significantly. Although filtering methods are simpler to implement, they are often criticized for selecting many redundant features and lack of considering interaction between the features (Sun, Goodison et al. 2007). To make the best use of groups of variables, wrapper models (Kohavi and John 1997) search for an optimal subset of features in the whole feature space and then evaluate that subset based on predictive models built using it. However, searching for an optimal subset of features is NP-hard (Amaldi and Kann 1998). As a result, most existing solutions depend on heuristic combinatorial search based on an objective function with a possible trade-off for optimality. Even so, it is difficult sometimes to define an appropriate objective function that can be optimized easily in high-dimensional data. Some clinicogenomic models (Sun, Goodison et al. 2007) aim at developing a method (e.g. I-RELEIF (Li and Sun 2006)) for handling this issue. However, both filtering techniques and wrapper models try to reduce dimensionality only on gene expression data and thus, they cannot find possible redundancy between the clinical and genomic data. Embedded models rather perform both feature subset selection and predictive learning in a single step. In clinicogenomic studies (Binder and Schumacher 2008), (Ma and Huang 2007), (Bovelstad, Nygard et al. 2009), (Boulesteix and Hothorn 2010), several regularization based predictive models have been developed with special modifications that impose two different penalties for the two types of data.

Some studies (Beane, Sebastiani et al. 2008), (Acharya, Hsu et al. 2008), (Berlingerio, Bonchi et al. 2009) rely on a previous study or prior biological knowledge to select the most appropriate genes for the particular disease they are investigating, rather than only depending on the dimensionality reduction techniques. All of these above mentioned feature selection techniques have their own advantages and disadvantages. Some clinicogenomic studies (Shedden, Taylor et al. 2008),(Bovelstad, Nygard et al. 2009) have assessed several dimensionality reduction methods to get the best classification results.

C. Discriminant models:

Discriminant models learn a discriminant function L = g(**x**) = $\mathbf{w}^T\mathbf{x}$ + w_0, where w is the coefficients for each variable of **x** as shown in Figure 1 for two-dimensional dataset **x** (**x** denotes the genomic variables here, but can also denote the clinical variables **c** and clinicogenomic variables **z**). LDA chooses the parameters **w** and w_0, such that the samples from the two classes are well-separated, maximizing the between class variances(Bishop and SpringerLink 2006).

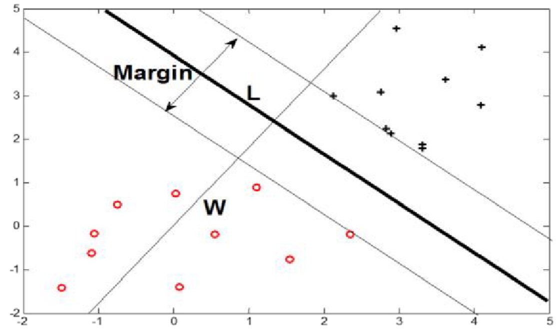

Figure 1: Discriminant models.

In addition to learning linear decision boundary, logistic regression (Duda, Hart et al. 2001),(Bishop and SpringerLink 2006) learns the posterior probability of the outcome variable by a logistic function **y**=sigmoid(g(**x**)). Logistic regression is a generalized linear model that summarizes the contributions of all predictors into a single variable, which is fed into a sigmoid transfer function to produce the final predicted probability of outcome event **y.** On the other hand, SVM tries to learn the decision boundary in such a way that it maximizes the separation between the two classes (measure by the soft margin).

Although logistic regression and LDA provide simpler discriminant models, they are typically confined in finding linear decision boundary only. Support vector machines (SVM) (Vapnik 2000) can circumvent this problem to learn more generalized non-linear decision boundary, by utilizing the power of kernel machines. The kernel machines first transfer the original feature space into higher dimensions by a non-linear mapping function and then, linear SVM is applied in that higher dimensional space. Thus, learning linear decision boundary in the higher dimensional space yields a non-linear decision boundary in the original space.

Tables and Figures:

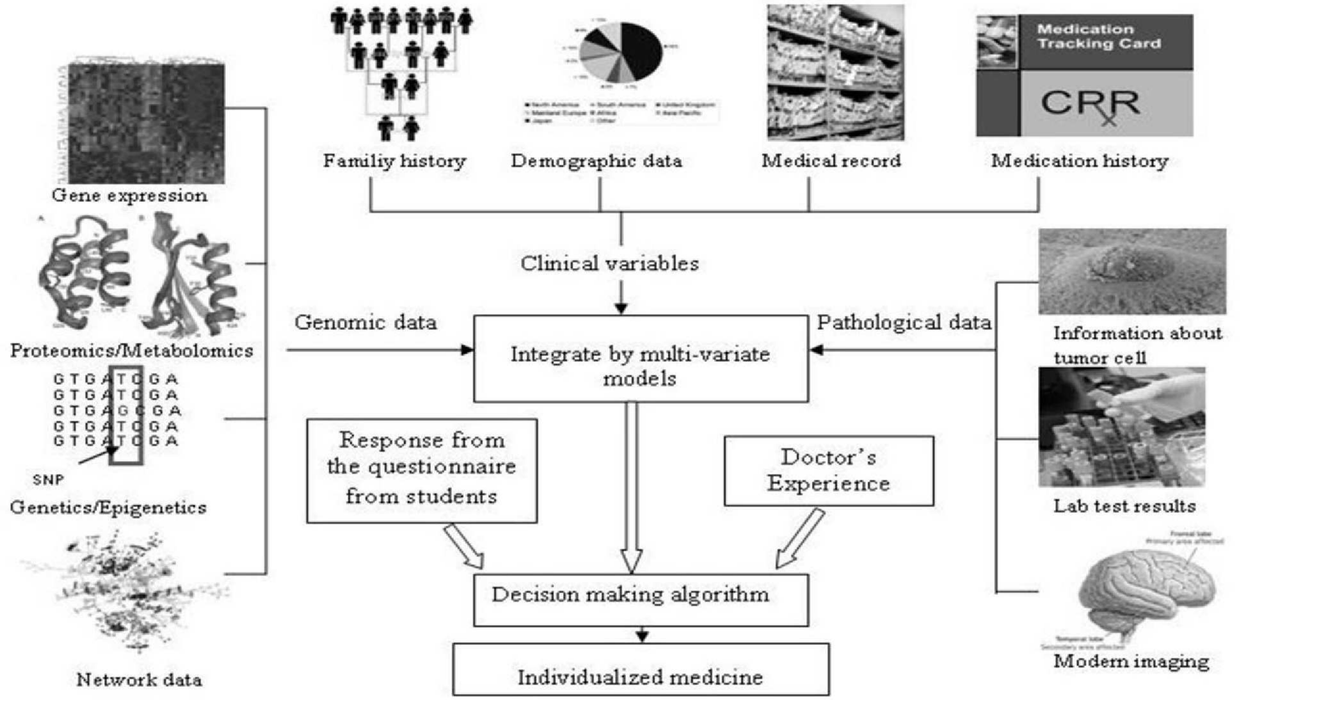

Figure 2: The integrative multivariate approach for clinical decision making by integrating multiple types of

Table 1: Related works of clinicogenomic studies

Studies	Challenges	Stage of integration	Ways to handle different dimensionalities	Disease heterogeneity	Predictive model	Added predictive value	Finding relationships	Validation	'Omic' Data	Clinical Data
Boulesteix et al 2011								√	√	√
Thomas et al. 2010							Interactions		√	Environmental
Correa et al. 2010							Correlations		√	√
Hamid et al. 2009	√	√			√				√	
Tsiliki et al. 2011	√				√		√		√	
Bebek et al. 2012							√		√	
Our review	√	√	√	√	√	√	√	√	√	√

Table 2: Taxonomy of different clinicogenomic models. Some branches are missing indicating no studies observed in that category. The studies marked as (M) means they are multi-site meta-analysis approach.

Main Categories		Early integration		Intermediate	Late
		Fullspace	Subspace	Fullspace	Fullspace
Explicit Dimensionality Reduction	Predictive modeling	*Regression:* -Li 2006 -Teschendorff et al. 06(M) -Shedden et al. 08(M) *Classification:* -Stephenson et al. 05 -Sun et al. 07 -Li. et al. 05 -Beane et al. 08 *Tree based method:* -Nevins et al. 03 -Pittman et al. 03 -Clarke et al. 08	Wang et al. 07 Berlingerio et al. 09 Schwarz et al. 09	Daemen et al. 07 Gevaert et al. 07	Campone et al. 08 Silvaha et al. 09 Futschik et al. 03
	Testing additional power	Tibshirani et al. 02 Hofling et al. 08 Boulesteix et al. 08	Acharya et al. 09		
Sparse Modeling	Predictive modeling	Binder et al. 08 Bovelstad et al. 09 Ma et al. 07			
	Testing additional power	Boulesteix et al. 2010			
	Finding relationships	Correa et al. (CCA, MCCA) Calhoun et al.(pICA, jiCA)			

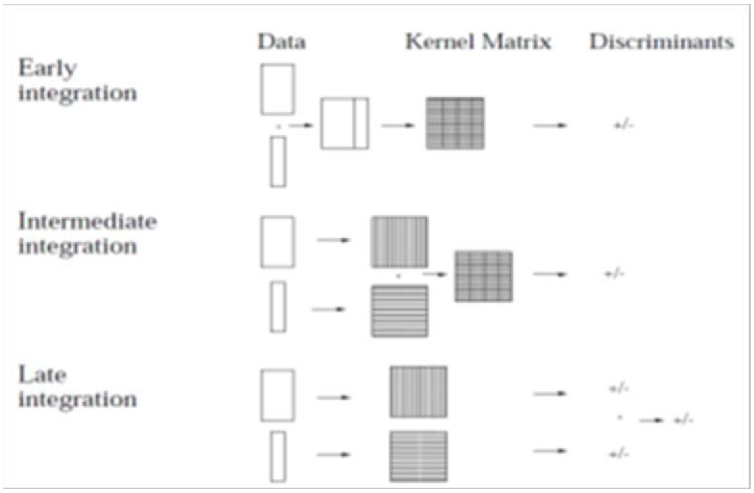

Figure 3: Pictorial representation of three types of integrations [taken from (Pavlidis, Weston et al. 2001)]

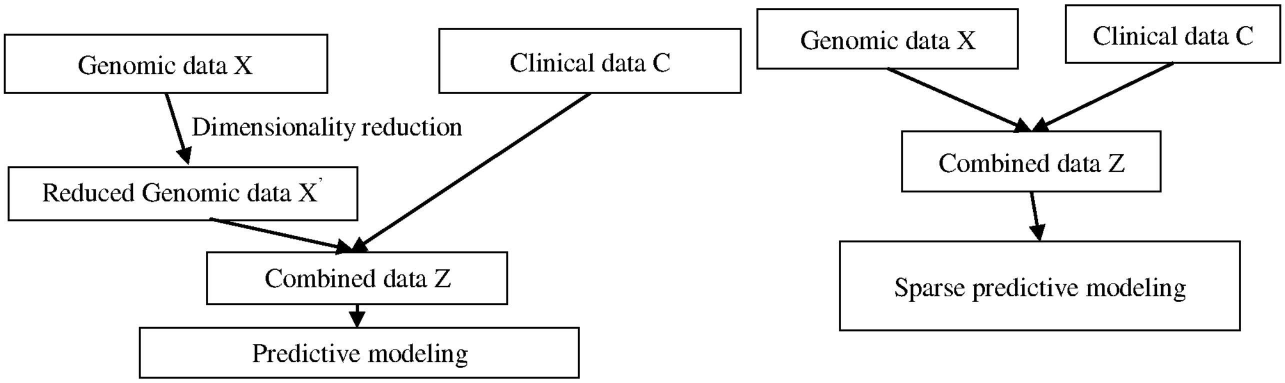

Figure 4: Two types of dimensionality reduction on the early integration process. a) The two-step approach where dimensionality reduction is performed on genomic data only. b) The one-step approach where dimensionality reduction is performed implicitly by sparse models.

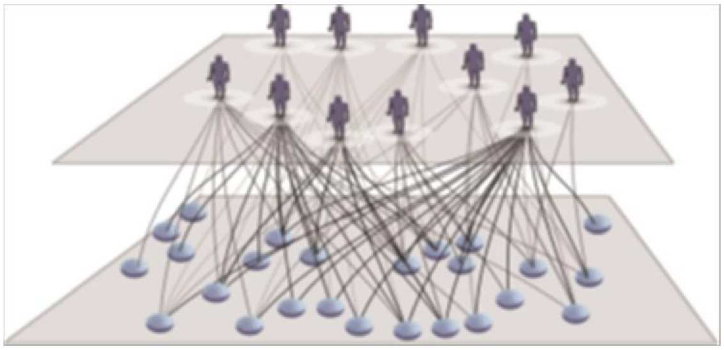

Figure 5: The complex disease network: the upper layer represents association between the patients and the lower layer represents the different candidate factors from laboratory test, genomic information, clinical factors and so on [(Schwarz, Leweke et al. 2009)]

Table 3: Summary of predictive clinicogenomic models

Clinicogenomic Study	**Stage of integration**	**Stage of dimensionality reduction**	**Full-space/subspace model**	**Dimensionality reduction technique**	**Predictive method**	**Clinical Endpoint**	Disease
Li 2006	Early	2-step	Fullspace	Two-step feature extraction: PCA, SIR	Cox hazard model	Survival time after chemotherapy	large-B-Cell lymphoma (DLBCL)
Stephenson et. al. 2005	Early	2-step	Fullspace	Ranking using statistical test	Logistic regression	Recurrance after Radical Prostatectomy (Binarized)	Prostate cancer
Sun et. Al., 2007	Early	2-step	Fullspace	Wrapper Model	Linear Discriminant analysis	Survival prediction (Two class)	Breast cancer
Li et. al, 2005	Early	2-step	Fullspace	Gene selection based on t-test	SVM	Response to platinum-based based Chemotherapy (Survival)	Ovarian cancer
Beane et al. 2008	Early	2-step	Fullspace	Ranking using Statistical test	Logistic regression	Development of metastasis after pathology	Lung cancer
Nevins et. al. 2003	Early	2-step	Fullspace	Feature creation by Clustering and PCA	Statistical tree	Survival prediction (Two class metastasis development)	Breast Cancer
Pittman et. al. 2004	Early	2-step	Fullspace	Feature creation by Clustering and PCA	Statistical tree	Survival prediction(Two class metastasis development)	Breast Cancer
Clarke et al. 2008	Early	2-step	Fullspace	Clustering based metagene	Statistical tree	Survival time after primary chemotherapy/disease relapse	Ovarian cancer
Cao et al. 2010	Early	2-step	Fullspace	Three dimensionality reduction methods	Mixture of Experts	Binary outcome	Breast cancer, Prostate cancer, Medulloblastomas
Binder et. al. 2008	Early	1-step	Fullspace	Regularization Techniques	Cox based prior model	Survival data	DLBCL
Ma et. al. 2007	Early	1-step	Fullspace	Statistical test and regularization	Penalized logistic and Cox regression	Binary class (metastasis), Survival analysis	Breast cancer, Follicular lymphoma
Bovelstad et. al. 2009	Early	1-step	Fullspace	Regularization method	Cox regression	Survival prediction	Breast cancer, DLBCL, Neuroblastoma
Berlingerio et al. 2009	Early	2-step	Subspace	Only HLA antigens/alleles corresponding to six loci are considered	Frequent Pattern Mining	Liver transplant VS. normal	Liver diseases leading to liver transplantation
Schwarz et. al. 2009	Early	2-step	Subspace	Domain guided	Network based framework	Case-control	Schizophrenia

Jana Silhava et. al. 2009	Late	2-step	Fullspace	Filtering	Logit, Bionomial boosting	Recurrance vs. Not recurrence	Breast Cancer
Campone et. al. 2008	Late	2-step	Fullspace	Filtering by univariant cox regression+PCA	Multivariate Cox regression analysis	Metastasis free survival	Breast cancer
Futschik et. al. 2003	Late	2-step	Fullspace	Filtering using statistical test	Bayesian network & ANN	Two class Survival after 5-yrs.	DLBCL
Daemen et. al. 2007	Intermediate	2-step	Fullspace	Ranking using statistical test	SVM	Metastasis	Breast cancer
Gevaert et al. 2006	Intermediate	2-step	Fullspace	Gene Filtering	Bayesian network	Metastasis	Breast cancer

Table 4: Clinicogenomic studies to assess additional prediction power of genomic features over clinical variables.

Clinicogenomic Study	**Stage of integration**	**Stage of dimensionality reduction**	**Full-space/subspace model**	**Dimensionality reduction technique**	**Predictive method**	**Clinical Endpoint**	**Disease**
Wang et. al. 2007	Early	2-step	Fullspace	Stepwise logistic regression for gene selection	SVM/SLD/ KNN	Recurrent vs. Non-recurrent	Human Hepatocellular Carcinoma)
Tibshirani et al. 2002	Early	2-step	Fullspace	Filtering approach based on p-value of fold change	Logistic regression	Binary class (metastasis vs. normal)	Breast cancer
Hofling et al. 2008	Early	2-step	Fullspace	Filtering approach based on p-value of fold change	Logistic regression	Binary class (metastasis vs. nomral)	Breast cancer
Boulesteix et al. 2008	Early	2-step	Fullspace	Supervised feature extraction, PLS	Tree based method	Binary class (metastasis vs. nomral)	Breast and Colorectal cancer
Boulesteix et al. 2010	Early	1-step	Fullspace	Regularization based technique	Logistic regression with boosting	Binary class (metastasis vs. normal & remission vs. no-remission)	Breast cancer, Acute Lymphoblastic Leucemia (ALL)

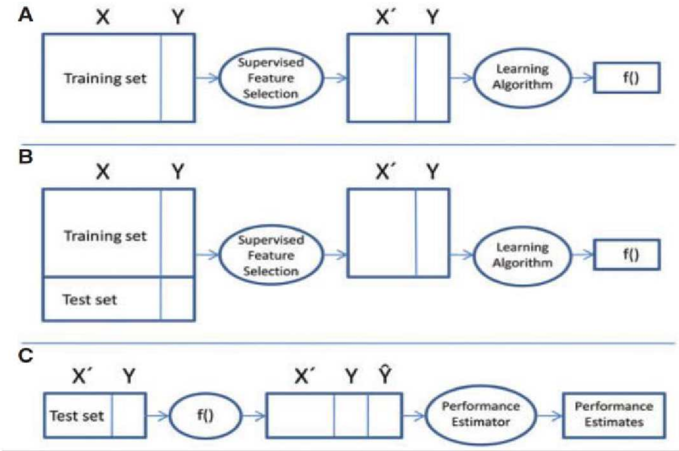

Figure 6: The right way to perform the supervised feature selection [(Smialowski, Frishman et al. 2010)] is step A followed by step C, where most of the methods follow step B that yields biased result for classification method.

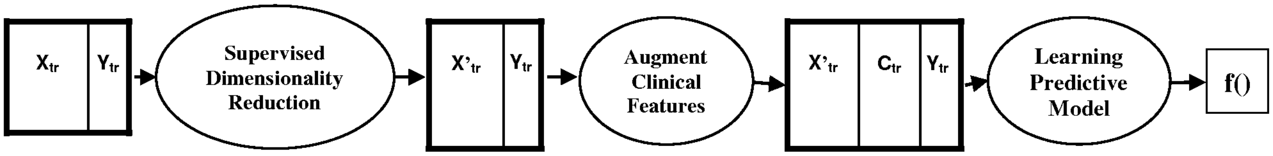
X_tr
Y_tr
Supervised Dimensionality Reduction
X'_tr
Y_tr
Augment Clinical Features
X'_tr
C_tr
Y_tr
Learning Predictive Model
f()

Figure 7: The training phase for clinicogenomic model (corresponding to Figure 6(a)).

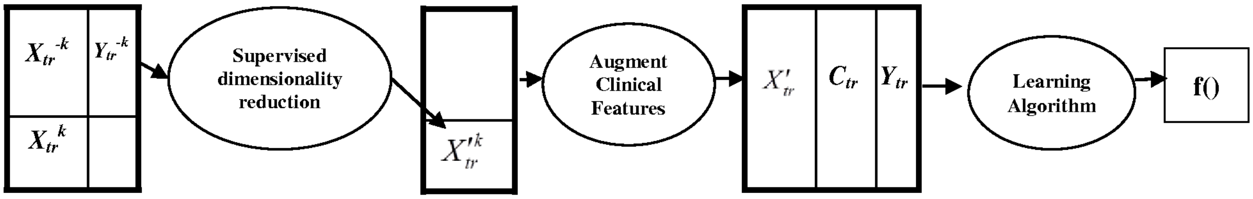

Figure 8: Schematic diagram of Pre-validation as suggested by Tibshirani et al. The first three phases are repeated here k times to get the full X' matrix. Here X_{tr}^{k} represents the k-th part of training set.

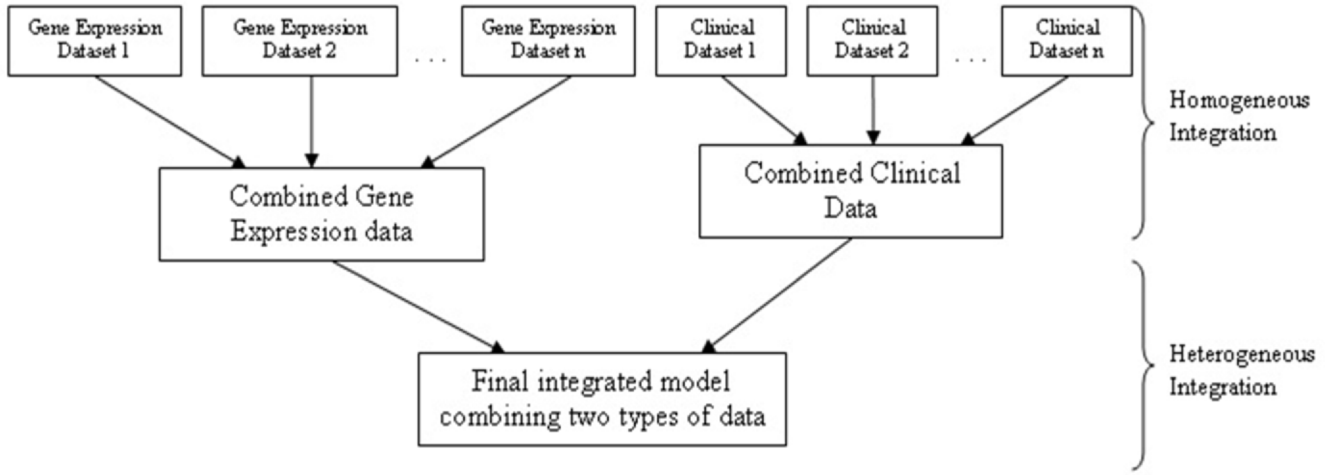
Gene Expression Dataset 1
Gene Expression Dataset 2
. . .
Gene Expression Dataset n
Clinical Dataset 1
Clinical Dataset 2
. . .
Clinical Dataset n
Combined Gene Expression data
Combined Clinical Data
Final integrated model combining two types of data
Homogeneous Integration
Heterogeneous Integration

Figure 9: The two stage integration of the multi-site clinicogenomic models.

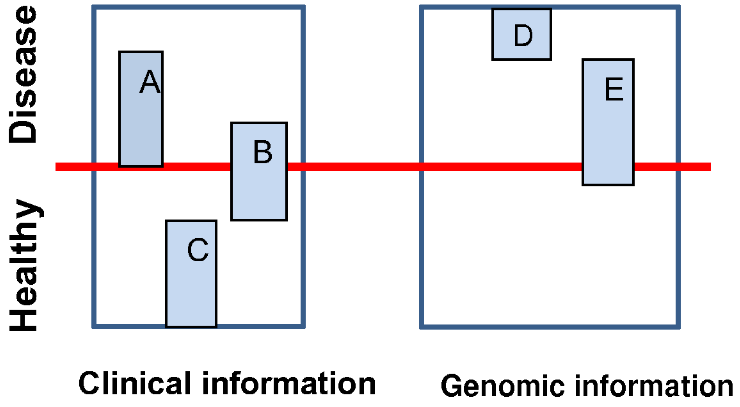

Figure 10: Subgroups of samples associated with several clinicogenomic markers. Each of the marker groups such as A-E corresponds to a subgroup. Among these A, C, E, D are discriminative (in order of their discriminative power). A and E cover same group of patients, where rest of them cover different subgroups of samples.

Table 5: Summary of multisite clinicogenomic studies.

Study	Dimensionality reduction technique	Predictive method	Integration type	Testing additive performance of genomic variables	Clinical Endpoint	Disease
Acharya et. al. 2009	Gene Clustering and preselection based on prior knowledge	Hierarchical Clustering	Early/Semi-supervised	Yes	Relapse free survival (may be distant)	Breast Cancer
Shedden et al. 2008	Comparitive study among 8 dimensionality reduction techniques	Cox hazard model	Early	No	Survival data	Lung Cancer
Teschendorff et. al. 2006	Common genes across 6 datasets	Univariate Cox model	Early	No	survival vs. death/ development of metastatis	ER+ breast cancer