Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Métodos de aprendizado de máquina supervisionado para classificação aplicados a dados de microarranjo de DNA.

Aluna: Ana Carolina Alves Oliveira
Orientadora: Samara Flamini Kiihl

Campinas
Set/2020

Sumário

Resumo

A tecnologia de microarranjos de DNA permite que cientistas monitorem expressões de vários genes ao mesmo tempo, tendo sido amplamente empregada para descobrir genes associados a fenótipos de interesse. Para identificar genes com expressão gênica diferenciada, diversas metodologias de aprendizado de máquina podem ser executadas. Neste trabalho foram aplicados os algoritmos k-vizinhos mais próximos e análise de discriminante linear em um conjunto de dados reais com informações de expressão gênicas de pacientes com determinada doença e de um respectivo grupo controle. Dessa forma, foi possível classificar os indivíduos em grupos e avaliar a acurácia dessas predições

Palavras Chaves: Aprendizado de Máquina, Algoritmos Supervisionados, Classificação, Microarranjo de DNA.

1 Introdução

Ao longo dos últimos anos, a geração de uma grande quantidade de dados genômicos e proteômicos tem resultado em um acúmulo de dados biológicos que necessitam ser interpretados (Raza, 2010). As áreas de bioinformática e bioestatística vêm desenvolvendo algoritmos capazes de processar e analisar um volume cada vez maior de dados biológicos, muitos dos quais, provenientes de chips de microarranjo.

Os dados de microarranjo são oriundos de técnicas que consistem na análise do nível de luz fluorescente captada por meio de imagens de chips de microarranjo, sendo possível quantificar e qualificar expressões de genes de determinado organismo. A grande vantagem de tal técnica é investigação de milhares de transcritos de um tecido, de maneira simultânea (Giachetto, 2010).

Os algoritmos de aprendizado de máquina, muito utilizados nesse contexto, surgiram da necessidade de automação das análises de dados, tornando possível a construção de modelos em uma escala superior à capacidade humana. Quando aplicados a dados de microarranjo, podem potencializar novos desenvolvimentos e contribuir para o diagnóstico e tratamento de diversas doenças.

Diante desse cenário, o presente trabalho consiste no estudo e aplicação de algoritmos de aprendizado de máquina supervisionado para classificação, construindo classificadores que sejam capazes de auxiliar no processo de identificação de pacientes com determinada doença, nesse caso, o Transtorno do Espectro Autista (TEA).

2 Conceitos Biológicos

Os ácidos nucleicos são moléculas que contém informações genéticas responsáveis pela origem e desenvolvimento de todos seres vivos. Eles são formados por nucleotídeos que por sua vez são constituídos pelos compostos químicos: ácidos fosfórico, bases nitrogenadas e açúcares. Esse último componente caracteriza os ácidos nucleicos em dois grupos: desoxirribonucleico (DNA) e ribonucleico (RNA). Ambos estão representados esquematicamente na Figura 1.

O DNA é uma molécula contida no núcleo das células que armazena informações genéticas de um indivíduo. Ele é representado por pares de bases

nitrogenadas que formam uma fita dupla unidas por interações moleculares de pontes de hidrogênio. As quatro bases que o compõe são: adenina (A), timina (T), citosina (C) e guanina (G). Ele é dividido em pequenas unidades funcionais conhecidas como gene. O conjunto de genes em um indivíduo é denominado genoma.

O RNA, principal molécula ligada à produção de proteínas, também é composto pelas 4 bases, entretanto a timina é substituída pela uracila (U) e é representado por uma fita simples. Existem diferentes tipos de RNA que desempenham determinadas funções celulares: O RNA mensageiro, que porta consigo informações do DNA; o RNA transportador que é responsável pelo transporte de aminoácidos e o RNA ribossômico, que é o principal componente dos ribossomos. Todos estão diretamente relacionados.

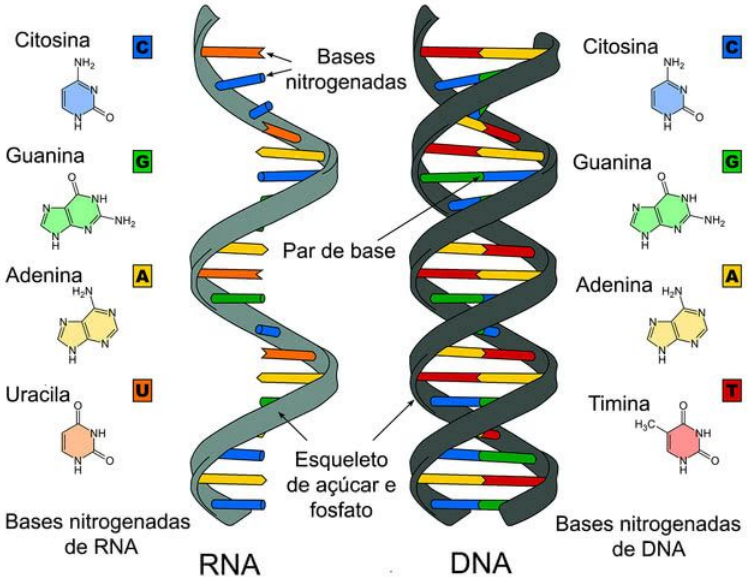

Figura 1: Representação das moléculas RNA e DNA
Fonte: Kaufman, 2020.

A partir da análise do DNA e RNA é possível detectar agentes infecciosos, mutações específicas que levam a proteínas com funções prejudicadas ou predisposição a doenças oncológicas, estabelecer prognóstico de doenças ou elencar o melhor tratamento a ser seguido especificamente para um paciente (Cabello, 2014).

3 Microarranjos de DNA

Os microarranjos ou chips de DNA (em alusão ao componente eletrônico que carrega milhões de transistores) são lâminas sólidas, nas quais fragmentos de DNA fita simples, denominados de sondas, são depositados e imobilizados de forma ordenada e em áreas específicas, chamadas de *spots* (Giachetto, 2010.). Para revestir os chips e atuar na detecção dos genes de interesse, são utilizados materiais bioquímicos como oligonucleotídios (moléculas simples de ácido nucleico com poucas bases em sua composição) e DNA complementar (cDNA - moléculas sintéticas de DNA derivadas do RNA mensageiro). Ambos também são conhecidos como sondas.

A técnica que utiliza chips de microarranjo se baseia na hibridização do DNA depositado na placa com um RNA mensageiro correspondente. O processo de hibridização ocorre quando há o pareamento de ácidos nucleicos e ele resultará em um acréscimo de luz fluorescente que poderá ser captado por um *scanner* de fluorescência. O processo de aquisição das imagens é de extrema importância, uma vez que, todas as análises realizadas dependerão delas. Após esse procedimento, a partir das imagens geradas é possível analisar a intensidade de luz fluorescente emitida utilizando métodos de bioinformática. Essa intensidade representa o quão expresso é determinado gene no indivíduo em estudo.

A preparação das lâminas que geram as imagens varia de acordo com protocolos de microarranjos, fabricantes, configurações do *scanner*, entre outros aspectos. Dessa forma há possibilidade de obter medidas de intensidades distorcidas que podem levar à conclusões erradas. Para contornar isso e garantir a confiabilidade das medias é realizado o pré-processamento dos dados, onde eles passam por correção de fundo, normalização e sumarização. Após esse procedimento os dados estão prontos para serem submetidos à análises estatísticas.

4 Aprendizado de Máquina

O Aprendizado de Máquina (AM), do termo inglês *Machine Learning* é uma área da inteligência artificial que surgiu no final do século XX. Pode ser definido como a capacidade de melhorar o desempenho na realização de alguma tarefa, por meio da experiência ou da descoberta de similaridade entre os dados (Mitchell, 1997). Seu objetivo é o desenvolvimento de técnicas com-

putacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática (Monard and Baranauskas, 2003). Diferente da programação de computadores tradicional, onde o programador insere os dados no programa e espera pelos resultados, no aprendizado de máquina, são informados os dados e os resultados e deseja-se obter o programa. Tal técnica possibilita tomadas de decisões e identificação de padrões pelas máquinas com o mínimo de interferência do homem. Duas das principais abordagens de AM são os algoritmos supervisionados e não-supervisionados.

Nos métodos de aprendizado de máquina supervisionado os algoritmos recebem um conjunto de variáveis independentes e precisam ser capazes de fazer previsões acerca de variáveis dependentes, que foram coletadas a partir das primeiras. A principal característica dos métodos supervisionados é que eles são treinados dispondo de pares de variáveis dependentes e independentes que são fornecidos pelo usuário. Baseado no treinamento dos algoritmos deseja-se predizer com bom desempenho novas observações, entender quais foram as variáveis independentes que afetam o resultado e mensurar a qualidade das previsões. Quando as variáveis dependentes são do tipo discretas, trabalha-se com algoritmos de classificação. Por outro lado, quando elas são contínuas são usados algoritmos de regressão (Monard and Baranauskas, 2003). Um exemplo para o primeiro caso é o estudo sobre um paciente sofrer ou não um ataque cardíaco. Enquanto para o segundo, o preço de ações após um determinado período.

No aprendizado não-supervisionado, são fornecidos pares de variáveis dependentes e independentes proveniente de dados coletados e esses são analisados com objetivo de tentar determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos (Cheeseman and Stutz, 1990). Esse tipo de algoritmo permite segmentar um conjunto de dados em grupos de acordo com características similares, encontrar inconformidades e identificar artefatos que ocorrem juntos nos dados. No presente trabalho, o foco será apenas no aprendizado supervisionado para classificação.

4.1 Algoritmos de Classificação

Os algoritmos de classificação são utilizados quando a variável dependente é do tipo discreta. Eles produzem um classificador capaz de generalizar as informações contidas no conjunto de treinamento, com a finalidade de classificar, posteriormente, objetos cuja classe seja desconhecida.

Dos diversos métodos de classificação desenvolvidos, os mais famosos são: análise de discriminante linear, regressão logística, *naïve bayes*, árvore de decisão, k-vizinhos mais próximos, entre muitos outros que a cada dia são aprimorados.

A escolha por qual utilizar irá depender das suposições necessárias para aplicar cada um, como por exemplo, da quantidade de dados disponíveis, da capacidade de processamento e da memória da máquina. Nas próximas seções, dois métodos serão abordado de forma detalhada.

Uma forma de avaliar o melhor modelo a ser utilizado é pela validação cruzada. Nesse procedimento os dados são separados em k subconjuntos (dobras) diferentes e repete-se k experimentos de treino e avaliação, cada vez mantendo um subconjunto diferente para avaliação. Por fim, é feita a média das performances de validação dos k experimentos para obter uma estimação do erro de generalização e do quanto o modelo acertou (acurácia). É comum atribuir ao k valores entre 5 e 10, entretanto não é uma regra (Facure, 2017).

4.1.1 K-Vizinhos Mais Próximos

O método K-Vizinhos Mais Próximos (*K-Nearest Neighbor*, KNN) é uma técnica que considera a proximidade entre dados na realização de predições. Ele supõe que os dados similares tendem a estar concentrados na mesma região no espaço de dispersão dos dados. O algoritmo é bem conhecido tanto em métodos de classificação, como de regressão. No primeiro, ele classifica uma nova observação de acordo com a classe da maioria de seus k-vizinhos mais próximos. Já no segundo, o algoritmo estima o valor de uma nova observação de acordo com a média dos valores de seus k-vizinhos mais próximos (Friedman, Hastie and Tibshirani, 2001).

Para o caso de classificação, considere os dados de treinamento (x_i, y_i) para $i = 1, ..., n$ e $x_i \in R^n$, uma observação com preditores x_0 e uma medida de distância $d(x_0, .)$. Escolhe-se um valor k, via validação cruzada e o método de classificação por KNN identifica k observações nos dados de treinamento que estão mais próximos de x_0 de acordo com $d(x_0, .)$. Ou seja, para todos os $x_i's$ de treinamento, calcula-se as respectivas distâncias. Um típico exemplo é a distância euclidiana, dada por (1).

$$d(x_0, x_i) = \sqrt{\sum_{j=1}^{p}(x_0 - x_i)^2} \tag{1}$$

onde $i = 1, ..., n$.

Após o cálculo, as distâncias $d(x_0, x_i)$ são ordenadas e armazena-se k valores de y_i correspondentes às menores distâncias (Friedman, Hastie and Tibshirani, 2001).

No caso de regressão, considere $\phi_k(x_0)$ a vizinhança que contém os k-vizinhos de x_0. A estimativa de y_0 será dada pela média de todas as respostas no treinamento pertencentes à $\phi_k(x_0)$, exibida em (2):

$$\hat{y_0} = f(\hat{x_0}) = \frac{1}{k} \sum_{x_i \in \phi_k(x_0)} y_i \tag{2}$$

4.1.2 Análise de Discriminante Linear

O método Análise de Discriminante Linear (*Linear Discriminant Analysis*, LDA) consiste na separação de dois ou mais grupos/classes por meio da realização de combinações lineares das suas características. Dessa forma é possível reduzir a dimensão dos dados, quando este possui muitas variáveis, antes de aplicar algoritmos de classificação.

De posse de um conjunto de dados, o LDA usa-o para dividir o espaço da variável dependente em regiões. Tais regiões são rotuladas por classes e têm limites lineares (Hoare, 2017). Assim, o algoritmo é capaz de prever o grupo a qual pertence uma nova observação baseada na região em que ela se encontra. Portanto, observações dentro do mesmo local pertencem a mesma classe.

O método LDA tem como pressuposto que todas as variáveis dependentes tem distribuição normal multivariada e a matriz de variâncias e covariâncias tem comportamento homogêneo entre as classes. Ainda que na prática essas suposições possam não ser totalmente válidas, sendo próximas é suficiente para o algoritmo funcionar bem.

Para fazer a modelagem é necessário selecionar quais serão as variáveis explicativas/independentes (discriminantes) e então identificar a função discriminante, que será responsável pela delimitação do espaço pertencente a

determinado grupo. Para o caso do número de classes igual a dois, a função é definida como em (3).

$$z_n = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \tag{3}$$

onde z_n é a variável dependente correspondente as possíveis classes, $x_1, ..., x_n$ são as variáveis explicativas/independentes. Os coeficientes discriminantes, $\beta_0, ..., \beta_n$, são os parâmetros que se tem interesse de estimar.

Para construção da regra de classificação é utilizada a função de discriminante linear de Fisher. Ela consiste na combinação linear das variáveis independentes (características) e se caracteriza por produzir a melhor separação entre classes.

5 Aplicação em Dados Reais

5.1 Aprendizado de Máquina no Diagnóstico de TEA

O Transtorno do Espectro Autista (TEA) é um distúrbio do neurodesenvolvimento infantil caracterizado por dificuldades na interação social, comunicação e comportamentos repetitivos, podendo apresentar também sensibilidades sensoriais (Fontes, 2014). Segundo o Centro de Controle e Prevenção de Doenças, nos Estados Unidos, a maioria das crianças não são diagnosticadas com TEA antes dos 4 anos de idade. Como a intervenção comportamental intensiva iniciada quando ainda bebê, produz bons resultados no desenvolvimento das crianças, há uma necessidade científica de biomarcadores que auxiliem em diagnósticos confiáveis de TEA e que sejam detectados no início da vida.

Diante desse contexto, o objetivo deste trabalho é aplicar técnicas de aprendizado de máquina supervisionado, a fim de classificar indivíduos com TEA em comparação com um grupo controle, com base em perfis de expressão de genes no sangue. Os perfis de expressão gênica podem ser usados como biomarcadores de diagnóstico para TEA.

5.2 Materiais e Métodos

Foi utilizado o conjunto de dados de Kuwano *et al* (2011), disponível em domínio público na plataforma *Gene Expression Omnibus* (GEO), sob o número de série GSE26415. Os dados consistem de 84 amostras de sangue venoso de indivíduos distribuídos nos seguintes grupos: adultos com TEA

(21), mães saudáveis que tiveram filhos com TEA (21) e grupo controle pareado com mesma idade e sexo de ambos grupos anteriores (42). Além disso, há expressões de 19195 genes registradas. No presente trabalho operou-se sob os dados já pré-processados, informações adicionais podem ser verificadas em Kuwano *et al* (2011).

Todo desenvolvimento computacional foi realizado no *software* RStudio, a análise é reprodutível e está disponível no *GitHub*, através do *link*: *github.com /anacarolina-estat/Iniciacao-Cientifica-CNPq*. O *download* dos dados foi feito utilizando as bibliotecas *Biobase* e *GEOquery*. Os algoritmos de aprendizado de máquina foram aplicados dispondo, principalmente, do pacote *caret*. Ele possui ferramentas de divisão dos dados, pré-processamento, ajuste, entre outras. Foram escolhidos os métodos K-Vizinhos Mais Próximos e Análise de Discriminante Linear, ambos por serem clássicos e conhecidos e pela simples compreensão quando se está iniciando os estudos em aprendizado de máquina.
Para treinar o algoritmo os dados foram divididos aleatoriamente de forma que 75% correspondem a dados de treinamento e 25% de teste. A variável dependente que se tem interesse em fazer previsões é o grupo/classe em que o indivíduo pertence: adulto com TEA, mães saudáveis que tiveram filhos com TEA ou grupo controle.

5.3 Resultados

5.3.1 K-Vizinhos Mais Próximos

Usando o algoritmo K-Vizinhos Mais Próximos nos dados de treinamento, com validação cruzada utilizando 5 dobras, foram obtidos os resultados presentes na Tabela 1. Nela são exibidos as métricas acurácia (proporção de casos previstos corretamente) e coeficiente Kappa (proporção de observações concordadas pelo valor real e predito) para diferentes valores de k, encontrados via validação cruzada. A partir dos resultados, o algoritmo seleciona automaticamente o melhor valor para k, que nesse caso corresponde a $k = 3$.

Tabela 1: Métricas do algoritmo KNN nos dados de treinamento, com validação cruzada com 5 dobras.

k	Acurácia	Kappa
1	0.57	0.34
2	0.59	0.36
3	0.67	0.46
4	0.60	0.36
5	0.57	0.32

De posse do modelo treinado é possível fazer previsões com os dados de teste e montar a matriz de confusão, que retorna os valores reais e preditos pelo algoritmos (Tabela 2).

Tabela 2: Matriz de confusão do algoritmo KNN aplicado aos dados de teste.

Predição	Referência		
	Adultos com TEA	Mães saudáveis com filhos com TEA	Grupo controle
Adultos com TEA	2	1	3
Mães saudáveis com filhos com TEA	3	0	2
Grupo controle	2	2	6

Ainda é possível explorar um pouco mais os valores previstos pelo modelo. Para isso são observadas algumas métricas da matriz de confusão, além da acurácia: a sensibilidade e especificidade. A primeira mensura a proporção de casos positivos que foram identificados corretamente, enquanto a segunda, a proporção de casos negativos também identificados corretamente. Esses resultados estão apresentados nas Tabelas 3.

Tabela 3: Métricas do algoritmo KNN aplicado aos dados de teste.

Métrica	**Classe**		
	Adultos com TEA	Mães saudáveis com filhos com TEA	Grupo controle
Sensibilidade	0.29	0	0.55
Especificidade	0.71	0.72	0.60
Acurácia	0.50	0.36	0.57

5.3.2 Análise de Discriminante Linear

Para treinar o algoritmo LDA foram usados os dados processados deixando-os normalizados e considerando as 10 primeiras componentes principais, a fim de reduzir o número de preditores e tornar o modelo mais simples. Outros valores de componentes foram testados mas não melhoraram o desempenho algoritmo. As métricas de precisão (acurácia) e coeficiente Kappa são 0.57 e 0.31, respectivamente.
A comparação entre valores reais e previsões encontradas a partir da aplicação do modelo nos dados de teste é mostrada na Tabela 4. Além disso, na Tabela 5 é possível visualizar métricas importantes da matriz de confusão.

Tabela 4: Matriz de confusão do algoritmo LDA aplicado aos dados de teste.

Predição	**Referência**		
	Adultos com TEA	Mães saudáveis com filhos com TEA	Grupo controle
Adultos com TEA	3	0	2
Mães saudáveis com filhos com TEA	2	1	2
Grupo controle	2	2	7

Tabela 5: Métricas algoritmo LDA aplicado aos dados de teste.

Métrica	Classe		
	Adultos com TEA	Mães saudáveis com filhos com TEA	Grupo controle
Sensibilidade	0.43	0.33	0.64
Especificidade	0.86	0.78	0.60
Acurácia	0.64	0.56	0.62

6 Conclusões e Considerações Finais

O presente trabalho foi desenvolvido com o intuito de aplicar métodos de aprendizado de máquina supervisionado para classificação em um conjunto de dados de microarranjo de DNA. Foram escolhidos dados de expressão gênica relacionados ao autismo e os algoritmos k-vizinhos mais próximos e análise de discriminante linear.

Foi levantada a relevância das técnicas de microarranjos, uma vez que, ela permite coletar dados extremamente importantes para detecção de doenças. Aliado ao aprendizado de máquina se torna possível otimizar a precisão dos diagnósticos e o tempo no qual são realizados.

Infelizmente ambos os algoritmos executados não tiveram desempenhos tão satisfatórios quanto se esperava, não sendo possível classificar tão precisamente os indivíduos no seus respectivos grupos. Embora o KNN tenha revelado uma melhor acurácia e Kappa quando treinado, nota-se que ao ser aplicado ao conjunto de dados de teste, não foi o que obteve as melhores métricas, especificamente quando se trata da classificação de indivíduos na classe “mães saudáveis que tiveram filhos com TEA”.

Uma possibilidade para melhoria no desempenho do modelo e obtenção de uma melhor discriminação dos grupos é realizar um filtro de gene. Esse processo consiste em analisar e selecionar genes que sejam mais expressivos entre os demais e posteriormente aplicar os algoritmos de aprendizado de máquina, levando em consideração apenas os que foram selecionados.

Bibliografia

- Cabello, G., 2014. Diagnóstico de doenças genéticas: Métodos de Rastreamento. Disponível em: *ghente.org/ciencia/genetica*. Acesso em: Junho de 2020
- Campos, R. e Tsang, R., 2008. Um estudo sobre processamento e análise de imagens de microarranjos de DNA. Disponível em: *www.cin.ufpe.br*. Acesso em: Dezembro de 2019
- Cheeseman, P. and J. Stutz. 1990. Bayesian classification (Autoclass): Theory and results advances in knowledge discovery and data mining.
- Christensen, DL, Baio, J, Van Naarden Braun, K, Bilder, D, Charles, J, and Constantino, JN (2016). Prevalence and characteristics of autism spectrum disorder among children aged 8 years–autism and developmental disabilities monitoring network, 11 sites, United States, 2012. MMWR Surveill Summ. 65, 1-23.
- Facure, M., 2017. Evaluating a Machine Learning Model.
- Fontes, A., 2014. Transtorno do Espectro Autista (TEA).
- Friedman, J., Hastie, T. and Tibshirani, R. 2001. The elements of statistical learning , 2th edition. Hair, J. F., R. L. Tatham, R. E. Anderson, and W. Black, 1998. Multivariate Data Analysis, 5th edition. Prentice Hall.
- Giachetto, P., 2010. A tecnologia de microarranjos na identificação de genes de interesse na bovinocultura.
- Gohlmann, H. and Talloen W., 2013. Gene Expression Studies. Using affymetrix microarray.
- Hoare J., 2017. Linear Discriminant Analysis in R: An Introduction. Disponível em: *displayr.com/linear-discriminant-analysis-in-r-an-introduction*. Acesso em: Julho de 2020
- Hosmer, D. W. and S. Lemeshow, 2000. Applied logistic regression. John Wiley and Sons.
- Kaufman, M., 2020. What, Exactly, Is A Virus?. Disponível em: *manyworlds.space/2020/03/23/what-exactly-is-a-virus/*. Acesso em: Agosto de 2020.
- Kuwano, Y, Kamio, Y, Kawai, T, Katsuura, S, Inada, N, and Takaki, A (2011). Autism-associated gene expression in peripheral leucocytes commonly observed between subjects with autism and healthy women having autistic children. PLoS One. 6, e24723.
- Mitchell, T. M. 1997. Machine Learning. New York: McGraw-Hill.
- Monard, M. C. and J. A. Baranauskas, 2003. Conceitos sobre aprendizado de máquina. In Sistemas Inteligentes Fundamentos e Aplicações, Pp. 89–114.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Raza, K. 2010. Application of data mining in bioinformatics. Indian Journal of Computer Science and Engineering.

- Warren, Z, McPheeters, ML, Sathe, N, Foss-Feig, JH, Glasser, A, and Veenstra-Vanderweele, J (2011). A systematic review of early intensive intervention for autism spectrum disorders. Pediatrics. 127, e1303-e1311.