Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Kernels como Ferramenta de Prognóstico da Leucemia Linfoide Aguda B-Derivada

Angie P. Soler[1], Marcos Eduardo Valle[2]
Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brasil.

Resumo. O repertório do receptor de células T (TCR) exerce papel central na resposta imunológica, sendo a sua diversidade um indicador relevante do estado do sistema imune. No contexto da leucemia, alterações nesse repertório podem refletir tanto a progressão da doença quanto a resposta ao tratamento. A complexidade do repertório de TCR torna a estimativa de diversidade um problema essencial, embora computacionalmente desafiador. Medidas tradicionais, como a entropia de Shannon e os números de Hill, capturam aspectos globais da distribuição, mas podem apresentar limitações diante da variabilidade dos dados imunológicos. Nesse cenário, métodos baseados em kernels têm se destacado como alternativas mais robustas. Neste trabalho, aplica-se o teste de Discrepância Média Máxima (MMD), em conjunto com um método de kernel, para comparar as distribuições dos repertórios de TCR de pacientes com leucemia em diferentes estágios do tratamento. Em contraste com abordagens clássicas, a metodologia proposta introduz um kernel ponderado por frequência, capaz de capturar recorrências nas sequências e de adaptar a estatística MMD às particularidades dos dados imunológicos. Com base em uma avaliação empírica da similaridade entre sequências, verifica-se que a abordagem empregada aprimora a distinção entre perfis leucêmicos e não leucêmicos, contribuindo para a identificação de padrões imunológicos associados à progressão da doença.

Palavras-chave. Leucemia linfoide aguda, discrepância média máxima, repertório do receptor de células T.

1 Introdução

Estudos recentes mostraram que a caracterização da diversidade do repertório imune pode ser utilizada para monitorar a resposta ao tratamento em pacientes com leucemia linfoide aguda (LLA) [5, 12]. A quantificação dessa diversidade é tradicionalmente realizada por meio de índices, como a entropia de Shannon [11] e os números de Hill [3], que medem a riqueza e a uniformidade. No entanto, essas abordagens frequentemente ignoram relações estruturais entre sequências, o que pode limitar a interpretação biológica dos dados [1, 4, 7].

A incorporação de métodos baseados em kernel tem se mostrado promissora na comparação de distribuições de sequências biológicas [8]. Em particular, o teste de Discrepância Média Máxima (MMD) tem sido aplicado na análise de diferenças entre distribuições de dados biológicos [6]. O MMD permite comparar diretamente distribuições de repertórios imunes projetando as sequências em um espaço de Hilbert e calculando uma medida de distância entre elas.

Neste estudo, explora-se a aplicação da MMD para quantificar a diversidade do repertório do receptor de células T em pacientes com LLA. Para isso, é empregada uma combinação de kernels modificados, integrando o kernel de espectro [8] com um kernel gaussiano adaptado por frequência, com o objetivo de aprimorar a precisão na classificação dos estados leucêmicos dos pacientes.

[1] angiesoler@ime.unicamp.br
[2] valle@ime.unicamp.br

2 Sequências Biológicas

Inicialmente, apresentam-se alguns aspectos biológicos relevantes para o estudo. Conforme discutido na introdução, os índices de diversidade convencionais consideram apenas a frequência das sequências, o que pode limitar a interpretação biológica. Para obter uma estimativa mais realista da diversidade, torna-se essencial definir uma medida de similaridade — como a proposta em [8] — que vá além da simples contagem de ocorrências, permitindo capturar relações estruturais entre as sequências.

Definição 2.1. *A* **sequência de nucleotídeos** *do DNA é uma sequência de caracteres definida em um alfabeto $\mathcal{A}_N = \{A, C, G, T\}$. As letras A, C, G e T, representam os nucleotídeos, respectivamente, adenina, citosina, guanina e timina.*

Definição 2.2. *A* **frequência** *da sequência biológica s_i num repertório $\mathcal{R}$ é*

$$p_i = \frac{\mathfrak{o}(s_i)}{\text{Número total de sequências}}, \tag{1}$$

com $\mathfrak{o}(s_i) :=$ número de ocorrências de s_i em $\mathcal{R}$.

A doença residual mínima (MRD) refere-se à presença de células leucêmicas remanescentes em níveis indetectáveis por avaliação morfológica convencional da medula óssea. Sua análise é essencial para monitorar a resposta ao tratamento e prever o risco de recaída.

2.1 Quantificação das Sequências Biológicas

Definição 2.3. *Um* ***k-mer*** *é uma subsequência de comprimento k contida em uma sequência biológica.*

Definição 2.4. *Dado $k \geq 1$, o* **espectro *k*** *de uma sequência biológica s é o conjunto de todos os $k-$mers de comprimento k que ela contém. O conjunto é denotado por $\sigma_k(s)$.*

Definição 2.5. *Para cada k-mer $\alpha \in \Sigma_k$, a* **coordenada indexada por** *α (ou seja, $\phi_\alpha(s)$) será o número de vezes em que α ocorre em uma sequência s. Isso fornece o* **mapa de características do espectro** *k, $\Phi_k : S \to \mathbb{R}^{l^k}$:*

$$\Phi_k(s) = (\phi_\alpha(s))_{\alpha \in \mathcal{A}^k},$$

com $\phi_\alpha : S \to \mathbb{R}$, $\phi_\alpha(s) =$ número de ocorrências de α em x, $\mathcal{A}$ o alfabeto de tamanho $|\mathcal{A}| = l$ e $\mathcal{A}^k = \underbrace{\mathcal{A} \times \cdots \times \mathcal{A}}_{k-\text{vezes}}$.

Definição 2.6. *O* **kernel do espectro *k*** *para duas sequências s_1 e s_2 é dado pelo produto interno no espaço de características:*

$$K_k(s_1, s_2) := \langle \Phi_k(s_1), \Phi_k(s_2) \rangle. \tag{2}$$

Neste estudo, busca-se um kernel que apresente maior capacidade de generalização, seja computacionalmente eficiente e possua maior suavidade — especialmente em razão da natureza discretizada da teoria empregada. Conforme discutido em [2], o kernel na Definição 2.6 apresenta limitações significativas: não generaliza bem para dados não vistos e falha em capturar similaridades parciais. Em particular, duas sequências com pequenas variações podem ser biologicamente muito semelhantes, mas esse kernel tradicional as considera completamente distintas. Para superar

DOI: 10.5540/03.2026.012.01.0301

essa limitação, adota-se o kernel modificado proposto em [2], que aprimora a capacidade de detecção de semelhanças estruturais e oferece maior flexibilidade na modelagem dos dados. O kernel modificado é dado por

$$\widehat{K}_k(s_1, s_2) = \langle \Phi_k(s_1), \Phi_k(s_2) \rangle \cdot \exp(-\gamma) + \lambda \cdot \exp\left(-\frac{\|\Phi_k(s_1) - \Phi_k(s_2)\|^2}{2\sigma^2} \right), \tag{3}$$

com $\lambda, \gamma, \sigma > 0$.

A seguir, definem-se os elementos básicos da teoria necessários à aplicação e à avaliação da MMD.

Definição 2.7. *Sejam S_i o* **conjunto de sequências biológicas da amostra** *i e S o* **conjunto de todas as sequências biológicas** *na base de dados, ou seja, a união dos S_i.*

Proposição 2.1. *Seja*

$$\eta_i(B) := \sum_{s_r \in B} \frac{\mathfrak{o}_i(s_r)}{N_i}, \tag{4}$$

para todo $B \subseteq S_i$, com $\mathfrak{o}_i$ o número de ocorrências da sequência $s_r \in S_i$ e $N_i = \sum_{l=1}^{|S_i|} \mathfrak{o}_i(s_l)$ o número total de sequências em S_i. A tripla $(S_i, \wp(S_i), \eta_i)$ é um espaço de medida.

Proposição 2.2. *Sejam $(S_i, \wp(S_i), \eta_i)$ um espaço de medida, f uma função mensurável, x uma variável aleatória com distribuição dada por η_i e $s_1, \ldots, s_n$ observações independentes com distribuição dada por η_i. A estimativa empírica de f é*

$$\mathbf{E}[f(x)] = \frac{1}{N_i} \sum_{j=1}^{n} f(s_j) \cdot \mathfrak{o}_i(s_j) \tag{5}$$

com N_i o número total de sequências em S_i.

3 MMD para Sequências Biológicas

A seguir, estabelece-se a base teórica do teste estatístico empregado. O estatístico do teste proposto por Gretton et al. [6] é definido como o maior valor absoluto da diferença entre as expectativas de funções pertencentes à bola unitária de um espaço de Hilbert reproduzível (RKHS). Essa formulação permite quantificar discrepâncias entre distribuições de forma robusta, sem a necessidade de suposições paramétricas. Além disso, pela Proposição 2.11 de [9], que se fundamenta no Teorema de Mercer, sabe-se que existe um mapeamento Φ para um espaço de Hilbert reproduzível ampliado $\widehat{\mathcal{H}}$, no qual o kernel transformado $\widehat{K}_k$ pode ser interpretado como um produto interno nesse espaço, ou seja,

$$\langle \Phi(x), \Phi(x') \rangle = \widehat{K}_k(x, x').$$

Essa propriedade permite reformular o teste estatístico no arcabouço de kernels, o que permite uma análise eficiente da discrepância entre distribuições.

No que se segue, $\widehat{\mathcal{H}}$ será o RKHS adotado. Sem perda de generalidade, assume-se que $\mathcal{F}$ é o conjunto de funções contidas na bola unitária de $\widehat{\mathcal{H}}$.

Definição 3.1. *Sejam $\mathcal{F}$ uma classe de funções $f : S \to \mathbb{R}$, η_1 e η_2 medidas de probabilidade, e $\mathfrak{o}_1$ e $\mathfrak{o}_2$ variáveis aleatórias com distribuições p_1 e p_2 dadas por η_1 e η_2, respectivamente. Considere*

amostras i.i.d. $\mathrm{S}_1 = \{s_1, \ldots, s_m\}$ *e* $\mathrm{S}_2 = \{\mathscr{s}_1, \ldots, \mathscr{s}_n\}$*, com distribuições* p_1 *e* p_2*, respectivamente. A* $\mathbf{MMD^2}$ *é definida para este caso através da equação*

$$\begin{aligned}\mathrm{MMD}^2[\mathcal{F}, S_1, S_2] =& \frac{1}{M(M-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \mathfrak{o}_1(s_i)\mathfrak{o}_1(s_j) \cdot \widehat{K}_k(s_i, s_j) \\ &+ \frac{1}{N(N-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathfrak{o}_2(\mathscr{s}_i)\mathfrak{o}_2(\mathscr{s}_j) \cdot \widehat{K}_k(\mathscr{s}_i, \mathscr{s}_j) \\ &- \frac{2}{MN} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathfrak{o}_1(s_i)\mathfrak{o}_2(\mathscr{s}_j) \cdot \widehat{K}_k(s_i, \mathscr{s}_j),\end{aligned} \tag{6}$$

em que N *e* M *representam os números de sequências totais considerando ocorrências, respectivamente, e* $\mathfrak{o}(\cdot)$ *é a função que conta o número de ocorrências.*

O estatístico MMD permite testar se duas distribuições p_1 e p_2 são distintas. Conforme o Teorema 5 do artigo [6], o MMD é zero se, e somente se, $p_1 = p_2$, ou seja, quando as distribuições são idênticas. O estatístico definido em (6) será utilizado como base para o teste de hipótese, possibilitando avaliar se as sequências do paciente diferem de forma estatisticamente significativa.

No contexto deste estudo, p_1 e p_2 representam as distribuições das sequências biológicas de um paciente em diferentes momentos: antes e depois do tratamento. Portanto, um valor elevado de MMD indica que a distribuição sofreu alteração, sugerindo uma redução na presença de células leucêmicas. Com efeito, um MMD significativo indica uma mudança na diversidade celular, associada a menor probabilidade de persistência da leucemia. Por outro lado, a ausência de diferença significativa será interpretada como indicativo de alta probabilidade de que o paciente permaneça leucêmico após o tratamento.

Para determinar a significância do MMD, utiliza-se o Teorema 3.1, cujos detalhes estão apresentados na Seção 5.5.2 do livro [10]. Especificamente, sejam $x_1, x_2, \ldots$ observações com distribuição dada por p, e seja K um kernel simétrico. Definem-se ζ_0, ζ_1 e ζ_2 como

$$\zeta_0 = \mathbf{E}_{x,x'} K(x, x'), \tag{7a}$$

$$\zeta_1(x_i) = \mathbf{E}_x K(x_i, x) - \zeta_0, \tag{7b}$$

$$\zeta_2(x_i, x_j) = K(x_i, x_j) - \zeta_1(x_i) - \zeta_1(x_j) - \zeta_0. \tag{7c}$$

Assim, ζ_1 corresponde à projeção de primeira ordem (isto é, à contribuição de cada amostra). Esse termo mede quanto a esperança de K varia quando condicionada a x_i. Já ζ_2 representa a projeção de segunda ordem (isto é, a interação entre pares de amostras). O seguinte teorema estabelece o comportamento assintótico do estatístico que estamos considerando.

Teorema 3.1. *Seja* K *um kernel simétrico e considere o estatístico* U_n *definido como*

$$U_n = \frac{1}{m(m-1)} \sum_{i \neq j} K(s_i, s_j). \tag{8}$$

Seja ainda

$$\theta = \mathbf{E}[K(s_i, s_j)], \tag{9}$$

a esperança da função de kernel sob a distribuição conjunta das amostras. Suponha que $\mathbf{E}[K^2] < \infty$ *e que a decomposição de Hoeffding do estimador* U_n *satisfaz* $\zeta_1 = 0$ *e* $\zeta_2 > 0$*. Então*

$$n(U_n - \theta) \xrightarrow{d} \frac{m(m-1)}{2} Y, \tag{10}$$

em que Y é uma variável aleatória da forma

$$Y = \sum_{l=1}^{r} \lambda_l(\chi^2_{1l} - 1), \tag{11}$$

com $\chi^2_{11}, \dots \chi^2_{1r}$ variáveis independentes e λ_l os autovalores associados à decomposição espectral do kernel K_k.

Com essa fundamentação, modifica-se o Teorema 12 do artigo [6] para estabelecer a distribuição nula do teste de hipótese baseado no MMD, o que permite avaliar se houve uma alteração significativa na diversidade celular e, consequentemente, inferir a probabilidade de persistência da leucemia.

Teorema 3.2. *Seja $\widetilde{K}_k(s_i, s_j)$ o kernel entre o mapa de características do qual a imersão média de μ_{η_1} tem sido subtraída,*

$$\begin{aligned}\widetilde{K}_k(s_i, s_j) :=& \langle \phi(s_i) - \mu_{\eta_1}, \phi(s_j) - \mu_{\eta_1} \rangle_{\mathcal{H}} \\ =& K_k(s_i, s_j) - \mathbf{E}_{\mathrm{S}_1^i} K_k(s_i, \mathrm{S}_1^i) - \mathbf{E}_{\mathrm{S}_1^i} K_k(\mathrm{S}_1^i, s_j) + \mathbf{E}_{\mathrm{S}_1^i, \mathrm{S}'} K_k(\mathrm{S}_1^i, \mathrm{S}'),\end{aligned} \tag{12}$$

com S' sendo uma cópia independente de S_1^i dada por p_1. Suponha que $t = m+n$, $\lim_{m,n\to\infty} m/t \to \rho_1$ e $\lim_{m,n\to\infty} n/t \to \rho_2 := (1 - \rho_1)$ para $0 < \rho_1 < 1$ fixo. Então, sobre a hipótese nula H_0, MMD_u^2 converge em distribuição a

$$tMMD_u^2[\mathcal{F}, \mathrm{S}_1, \mathrm{S}_2] \underset{D}{\longrightarrow} \sum_{l=1}^{r} \widetilde{\lambda}_l \left[(\rho_1^{-1/2} a_l - \rho_2^{-1/2} b_l)^2 - (\rho_1 \rho_2)^{-1} \right], \tag{13}$$

com $r = \mathrm{posto}(A)$, $A_{ij} = \widetilde{K}_k(s_i, s_j)$, $a_l \sim \mathcal{N}(0,1)$ e $b_l \sim \mathcal{N}(0,1)$ sendo sequências finitas de variáveis aleatórias gaussianas independentes, e os $\widetilde{\lambda}_i$ são os autovalores normalizados de $\widetilde{K}_k$.

4 Resultados e Discussão

Nesta seção, apresentam-se os resultados obtidos a partir da aplicação da teoria desenvolvida nas seções anteriores, o processo de obtenção dos dados conduzido pelo grupo de pesquisa do Centro Infantil Boldrini e uma análise preliminar dos estatísticos gerados.

A preparação e seleção das sequências biológicas de cada paciente foram realizadas conforme descrito na seção de métodos de [5]. O estudo inclui amostras de células da medula óssea de 76 pacientes com leucemia linfoide aguda B-derivada, analisadas no Centro Infantil Boldrini, em Campinas, São Paulo. Para a análise da MRD e a avaliação da frequência clonotípica de cada paciente, utilizam-se amostras pareadas coletadas no Dia 0 e no Dia 35 do tratamento.

Todas as amostras do Dia 0 são classificadas como leucêmicas e apresentam um alto índice de linfoblastos, o que possibilita a identificação das sequências de nucleotídeos associadas a esses clonótipos. As amostras do Dia 35, por sua vez, são empregadas na estimativa da MRD, além de fornecerem informações sobre a abundância das sequências clonotípicas ao longo do tratamento.

A fim de testar a hipótese de que a distribuição das sequências permaneceu inalterada ou sofreu mudanças significativas durante o tratamento, aplica-se o estatístico MMD, definido na equação (6), com nível de significância de 0.01, às sequências de cada paciente nos dois instantes de coleta, desconsiderando as sequências leucêmicas do Dia 35. Esse procedimento resulta em um valor de MMD por paciente, totalizando 76 observações. Além disso, avalia-se a relação entre os valores de

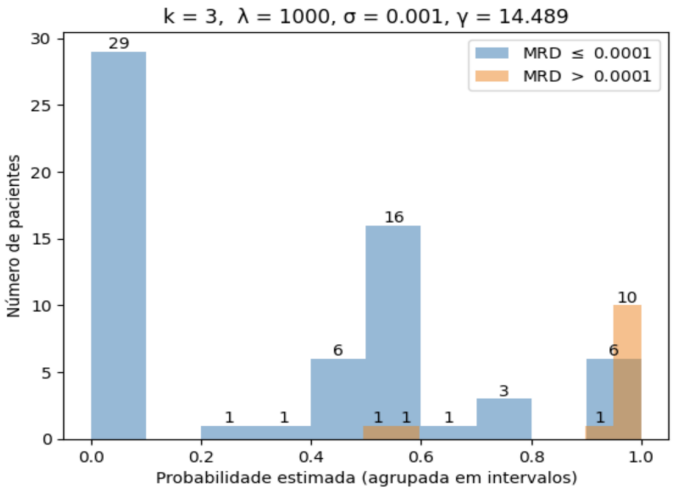

Figura 1: Probabilidade de permanência da leucemia após o tratamento, agrupada conforme o estado leucêmico determinado pela MRD. Fonte: Produzido pelos autores.

probabilidade calculados a partir da MMD e da MRD, o que permite a classificação dos pacientes de acordo com seu estado leucêmico no Dia 35, além de possibilitar, futuramente, a determinação de um limiar ótimo para essa discriminação, sem a necessidade do uso da MRD.

Utiliza-se a otimização bayesiana para selecionar os valores ótimos dos parâmetros $(\lambda, \sigma, \gamma, \alpha)$ que maximizam a correlação de Spearman entre os valores de MRD e os p-valores obtidos pela MMD em todos os conjuntos de sequências dos pacientes. No contexto da leucemia, pacientes com MRD ≤ 0.0001 são considerados em remissão, enquanto aqueles com MRD > 0.0001 ainda apresentam sinais da doença.

A Figura 1 revela divergências entre as classificações obtidas pela MRD e os valores de probabilidade calculados a partir da MMD. Alguns pacientes classificados como não leucêmicos pela MRD apresentam probabilidade de 1 de recaída segundo a MMD, enquanto outros, classificados como leucêmicos, apresentam probabilidade próxima de 0.5 de ainda estarem doentes.

Ao utilizar a MRD como referência tanto para a classificação dos pacientes quanto para a determinação dos parâmetros ótimos da MMD, observam-se 8 pacientes incorretamente classificados, dos quais apenas 6 representariam um possível risco clínico.

5 Conclusão e Considerações Finais

Com o auxílio do estatístico MMD, proposto em [6], e do kernel baseado em k-mers apresentado em [8] e [2], foi possível extrair informações quantitativas das sequências biológicas analisadas. Essa abordagem permitiu manipular os dados de forma mais eficiente, facilitando sua interpretação e aplicação. Em particular, os resultados do teste de hipótese forneceram estimativas da probabilidade de permanência da leucemia nos pacientes ao longo do tratamento.

Apesar dos avanços alcançados, observa-se que as probabilidades inferidas a partir do MMD não correspondem exatamente aos valores clínicos esperados, havendo casos em que a estimativa diverge do diagnóstico real. Para aprimorar a precisão do modelo, serão necessários novos experimentos, incluindo a análise da influência dos parâmetros k, λ, σ e γ sobre o desempenho do método. Ainda assim, os resultados obtidos até o momento indicam que a abordagem proposta constitui uma boa aproximação inicial, abrindo caminho para futuras melhorias e validações.

Agradecimentos

Angie P. Soler agradece o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Código de Financiamento 001. Marcos Eduardo Valle agradece o apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo 2023/03368-0, e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo 315820/2021-7.

Referências

[1] B. Allen, M. Kon e Y. Bar-Yam. "A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats". Em: **Am. Nat.** 174 (2009), pp. 236–243. DOI: 10.1086/600101.

[2] A. N. Amin, E. N. Weinstein e D. S. Marks. "Biological sequence kernels with guaranteed flexibility". Em: **arXiv preprint arXiv:2304.03775** (2023).

[3] A. Chao, C. H. Chiu e L. Jost. "Phylogenetic diversity measures based on Hill numbers". Em: **Philos. Trans. R. Soc. B** 365 (2010), pp. 3599–3609. DOI: 10.1098/rstb.2010.0272.

[4] D. P. Faith. "Conservation evaluation and phylogenetic diversity". Em: **Biol. Conserv.** 61 (1992), pp. 1–10. DOI: 10.1016/0006-3207(92)91201-3.

[5] G. N. N. Giusti. "Sequenciamento de alto rendimento para quantificação de doença residual mínima em leucemia linfóide aguda". Acessado em: 2 Jul. 2024. Dissertação de mestrado. Universidade Estadual de Campinas, Instituto de Biologia, 2019. URL: https://hdl.handle.net/20.500.12733/1637195.

[6] A. Gretton et al. "A Kernel Two-Sample Test". Em: **Journal of Machine Learning Research** 13.25 (2012), pp. 723–773. DOI: 10.48550/arXiv.2004.1109.

[7] T. Leinster e C. Cobbold. "Measuring diversity: the importance of species similarity". Em: **Ecology** 93.3 (2012), pp. 477–489. DOI: 10.1890/10-2402.1.

[8] C. S. Leslie et al. "Mismatch string kernels for discriminative protein classification". Em: **Bioinformatics** 20.4 (2004), pp. 467–476. DOI: 10.1093/bioinformatics/btg431.

[9] B. Schölkopf e A. Smola. **Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond**. MIT Press, 2002. ISBN: 9780262256933.

[10] R. J. Serfling. **Approximation Theorems of Mathematical Statistics**. John Wiley & Sons, 1980. ISBN: 9780471024033.

[11] C. E. Shannon. "A Mathematical Theory of Communication". Em: **Bell System Technical Journal** 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[12] F. F. Silva e M. R. D. O. Latorre. "Sobrevida das leucemias linfoides agudas em crianças no Município de São Paulo, Brasil". Em: **Cad. Saúde Pública** 36.3 (2020). DOI: 0102-311X00008019.

