Genetics inMedicine | EDUCATION REPORT

Open

Optimizing genetics online resources for diverse readers

Jiyoo Chang, BA[1,2], Monica Penon-Portmann, MD[1,2] and Joseph T. Shieh, MD, PhD[1,2]

Purpose: Clear and accurate genetic information should be available to health-care consumers at an individualized level of comprehension. The objective of this study is to evaluate the complexity of common online resources and to simplify text content using automated text processing tools.

Methods: We extracted all text from Genetics Home Reference and MedlinePlus in bulk and analyzed content using natural language processing. We applied custom tools to improve the readability and compared readability before and after text optimization.

Results: Commonly used educational materials were more complex than the recommended reading level for the general public. Genetic health information entries from Genetics Home Reference ($n = 1279$) were written at a median 13.0 grade level. MedlinePlus entries, which are not exclusively genetic ($n = 1030$), had a median grade level of 7.7. When we optimized text for the 59 actionable conditions by prioritizing medical details using a standard structure, the average reading grade level improved.

Conclusion: Factors that increase complexity are long sentences and difficult words. Future strategies to reduce complexity include prioritizing relevant details and using more illustrations. Simplifying and providing standardized online health resources would benefit diverse consumers and promote inclusivity.

Genetics in Medicine (2020) 22:640–645; https://doi.org/10.1038/s41436-019-0695-7

Keywords: consumer health informatics; natural language processing; genomics; educational resources; infographics

INTRODUCTION

In the digital age, the Internet is a key source of information for many. With the expansion of information online, the use of the Internet for health information has been growing as well. In 2018, up to 89% of US adults were reported to be Internet users.[1] Broadly available and reliable health resources allow patients and providers to share trusted information and discuss medical management.

Despite the growing volume of online health information, its universal impact can be limited by the complexity of the information presented.[2] Prior studies show that educational materials for patients, across different medical fields, are more complex[3–5] than the average reading level in the United States, which is around 8th grade.[6] The National Institutes of Health (NIH) recommends that patient education material is written at a 7th and 8th grade level.[7] Optimizing comprehension of health information is important as it relates to health literacy and health outcomes. Numerous studies have shown that poor health literacy is associated with worse health outcomes.[8]

Patients seeking information on genetic conditions may have greater difficulty finding comprehensible information online. Genetic conditions are individually less common, and less information is readily available. In addition, genetic information can be technical and complex and can require understanding of underlying biological concepts. One study showed that patients viewed genetics as a "specialist, scientific subject."[9] Many patients, however, turn to the Internet to find content about their conditions.[10] Online health information is not always screened for accuracy, and there are not many measures to evaluate the quality of information online.[11] Thus, resources such as Genetics Home Reference and newborn health screening programs have provided reliable and trustworthy information for patients and general health-care providers. Appropriate communication of genetic information could lead to improved health promotion.

As genetics and genomics are increasingly applied to clinical care, there will be a growing demand for genetic health information from consumers. Thus, it is crucial to examine how genetic information is presented online for broad public consumption. To our knowledge, little is known about the complexity of common consumer-targeted information for genetic conditions. We aim (1) to assess readability of common web-based resources for medical genetics and (2) to improve content using automated text processing tools.

MATERIALS AND METHODS

Initial assessment of common resources

We assembled commonly used genetics web-based resources for text complexity analyses. Initially, we assessed web-based resources for phenylketonuria (PKU) from Genetics Home Reference[12] (GHR), MedlinePlus,[13] Genetic and Rare Disease Information Center,[14] and the National Center for

[1]Division of Medical Genetics, Department of Pediatrics, University of California San Francisco, San Francisco, CA, USA; [2]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA. Correspondence: Joseph T. Shieh (joseph.shieh2@ucsf.edu)

Submitted 5 April 2019; revised 25 October 2019; accepted: 28 October 2019
Published online: 26 November 2019

Biotechnology Information (NCBI) GeneReviews.[15] We analyzed text describing the condition from the four sources and compared the complexity level. We additionally assessed reading levels of text from patient support groups such as March of Dimes,[16] PKU.com,[17] Mayo Clinic,[18] National PKU Alliance,[19] and National PKU News.[20] Information about Li–Fraumeni syndrome was accessed from several NIH resources and from support groups including the Li–Fraumeni Syndrome Association,[21] Living LFS,[22] and the American Society of Clinical Oncology.[23]

Text complexity analysis of GHR and MedlinePlus

We have selected GHR and MedlinePlus for an in-depth analysis as they are reliable sources of patient-friendly information provided by the National Library of Medicine (NLM). GHR is a commonly used genetics health information source. MedlinePlus is written specifically for consumers. In addition, GHR contains a thorough repository of information on genetic conditions while MedlinePlus offers information about diseases that are not exclusively genetic conditions. We extracted bulk text from GHR (https://ghr.nlm.nih.gov/download/ghr-summaries.xml) and MedlinePlus (https://medlineplus.gov/xml.html). From each website, we downloaded XML of all entries available in October 2018 and converted it to plain text by automatically removing xml and html tags. Each entry was formatted and analyzed using a script in R software. We compared the readability of text for matching genetic conditions between the two resources ($n = 80$) prior to applying our methods to the whole data sets. For all texts, we calculated text complexity using a custom script and the koRpus package in R software (https://cran.r-project.org/web/packages/koRpus/index.html). The processed texts were systematically fed into the script to output the results of 24 different readability formulas including FOG, SMOG, FORCAST, ARI, Flesch–Kincaid, Dale–Chall, and Coleman–Liau. The formulas demonstrated general concordance, so we focused on two well-established methods: Flesch–Kincaid Grade Level and New Dale–Chall formula. For the Flesch–Kincaid Grade Level analyses, we assessed the number of words per sentence and the number of syllables per word to estimate the reading grade level. As the formula is based on polysyllabic words and long sentences, this could underestimate the reading difficulty of text. Thus, we also used the New Dale–Chall method, which calculates the grade level based the sentence length and also the number of "hard" words that are not in a list of 3000 familiar English words. We use natural language processing (NLP) tools to find and replace difficult words, generate new text templates, and pull text information from the NIH, NCBI, and National Library of Medicine resources. NLP methods allow exploration and computational analysis of text-based data and have various applications in biomedical data.[24]

Statistical analysis

We performed statistical tests using R. A paired z-test was used to compare conditions in GHR and MedlinePlus. Kruskal–Wallis rank sum test was used to compare the readability scores after text optimization with the original. Histograms and plots were made with R graphics and the ggplot2 package (http://ggplot2.org/).

RESULTS

We assessed several commonly used genetic health condition information pages for common genetic conditions. For example, web information for PKU is available from several resources: GHR, GeneReviews, Genetics and Rare Disease Information Center, and MedlinePlus. In addition, we assessed the reading levels of texts provided by five PKU patient support group and other consumer-targeted resources, which are displayed in Supplementary Fig. 1. Interestingly, the sources varied greatly in the reading grade level of the text (Fig. 1). The lowest reading grade-level content for PKU was provided by MedlinePlus (6.6 grade). GeneReviews, known as an in-depth resource for providers, was at a college reading level (15.8 grade). Only MedlinePlus was written in a way that met the 7th to 8th grade reading level recommended by NIH.

To compare the complexity of the information for more genetic conditions, we extracted text content for 80 matching genetic conditions that had entries in both GHR and MedlinePlus. When these matching genetic conditions between GHR and MedlinePlus were directly compared, GHR entries were 4.7 grade levels higher in complexity (Z-score, $p < 0.05$) (Fig. 2). Seventy-nine of 80 conditions had a lower reading grade level in MedlinePlus compared with GHR.

We then compared the readability of all entries in GHR with MedlinePlus. Genetic health information entries from GHR ($n = 1279$) had text that scored at a median 13.0

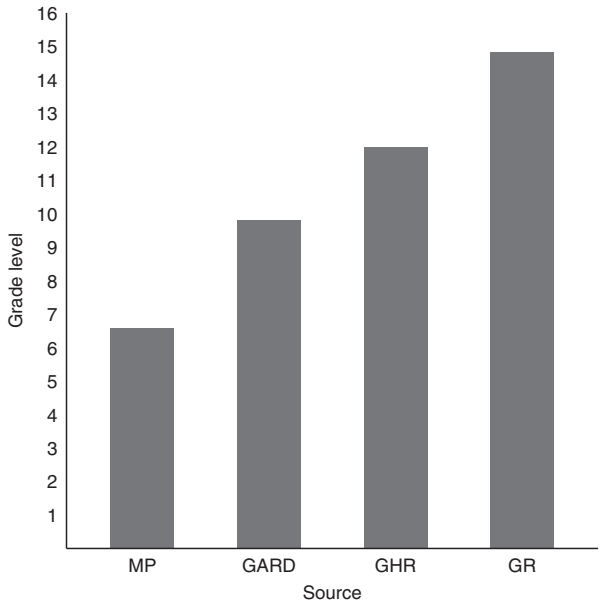

Fig. 1 Reading grade level of texts on phenylketonuria (PKU) from four online sources in ascending order of complexity. *GARD* Genetics and Rare Disease Information Center, *GHR* Genetics Home Reference, *GR* GeneReviews, *MP* MedlinePlus.

(SD = 1.7) grade reading level. In contrast, MedlinePlus entries ($n = 1030$), which are not exclusively genetic, had a median grade reading level of 7.7 (SD = 1.8) (Fig. **3a**). In terms of word complexity, 99% (1274/1279) of GHR entries were written at college level or above while 57% (587/1030) of MedlinePlus entries were written at college level or above, as estimated by the Dale–Chall method (Fig. **3b**). This demonstrates that commonly used patient educational materials are often more complex than the 7th to 8th grade level recommended for the general public. Since reading level may depend on several factors, we then examined why genetics text is complex and how to potentially simplify text.

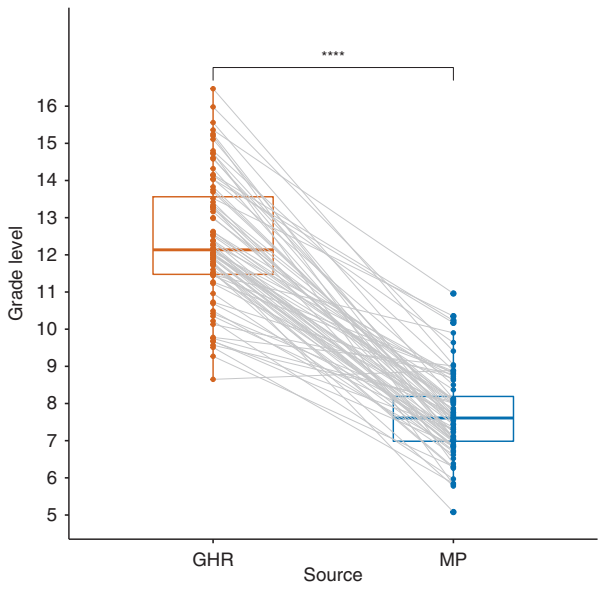

Fig. 2 Reading grade level of 80 matching genetic condition entries in Genetics Home Reference (GHR) (12.4 ± 1.7) and MedlinePlus (MP) (7.7 ± 1.1). Mean scores between the two sources are significantly different (Paired *z*-test, $p < 0.0001$).

Natural language processing

We applied NLP methods to improve readability for a set of conditions pertaining to the American College of Medical Genetics and Genomics (ACMG^{TM}) 59 conditions.[25] These are typically penetrant genetic conditions with actionable information that are reported as incidental or secondary findings in clinical genomic sequencing.[25] We applied NLP methods in a step-wise manner by first removing medical jargon and then replacing the complex condition name (steps 1 and 2 in Fig. 4). We compared readability scores before and after text optimization. When we programmatically processed the text with step 1 for the ACMG^{TM} conditions ($n = 28$), the average reading grade level moderately improved from 12.8 to 12.3. By replacing repeats of complex condition names (step 2), the score lowered to 11.6 (Kruskal–Wallis, $p < 0.05$) (Fig. 4). We also identified a set of long, complex words found in genetic resources that could be problematic for patients (Table S1). Many of these words are scientific terms that cannot be easily replaced or shortened. Since preliminary text processing methods (steps 1 and 2) only modestly lowered the reading grade level, we performed novel curation of informational resources to generate new educational content using NLP tools. Our text processing method (step 3) consisted of creating a new template for simplified genetic information by first bulk downloading of online health educational resources including GHR and MedlinePlus, as well as information from ClinGen actionability. For each condition, we extracted key medical details such as a short one-sentence description of condition, gene associated with the condition, risk associated with the condition, clinical actionability, and known inheritance, and integrated these details into a standardized template. This resulted in a simple structured summary of the condition (Fig. 5). After this step, the mean reading level of text decreased to 9.3 grade (Kruskal–Wallis, $p < 0.0001$) (Fig. 4). Structured summaries could be generated in a scalable fashion for consumer health information.

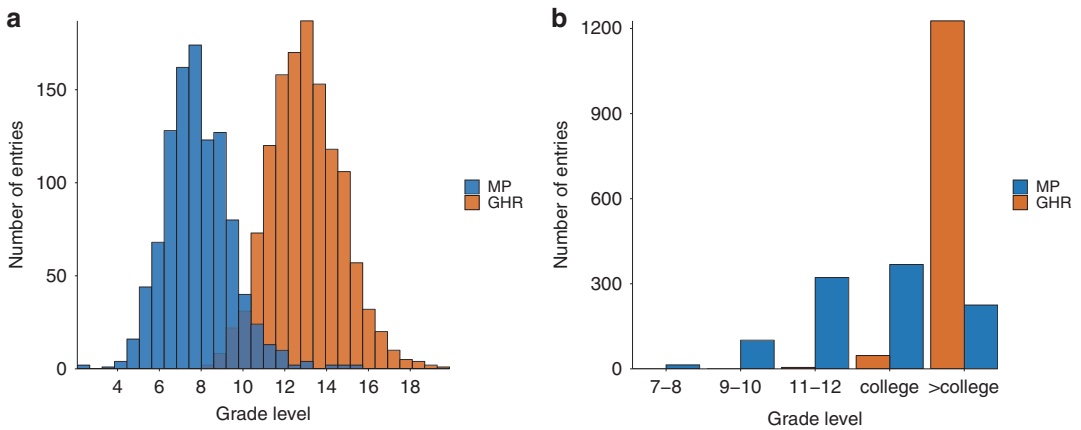

Fig. 3 Readability assessment of entries from Genetics Home Reference (GHR, $n = 1279$) and MedlinePlus (MP, $n = 1030$). (**a**) Distribution of readability scores for GHR (grade 13.0 ± 1.7) and MedlinePlus (grade 7.9 ± 1.8) using Flesch–Kincaid analyses (**b**) Word complexity assessment of GHR and MP using New Dale–Chall method.

DISCUSSION

Genetic information is reaching more individuals and their families with advances in sequencing technologies and increasing applications in clinical settings. For consumers to fully utilize genetic information and make appropriate health decisions, reliable information should be accessible and appropriate for the intended audience.

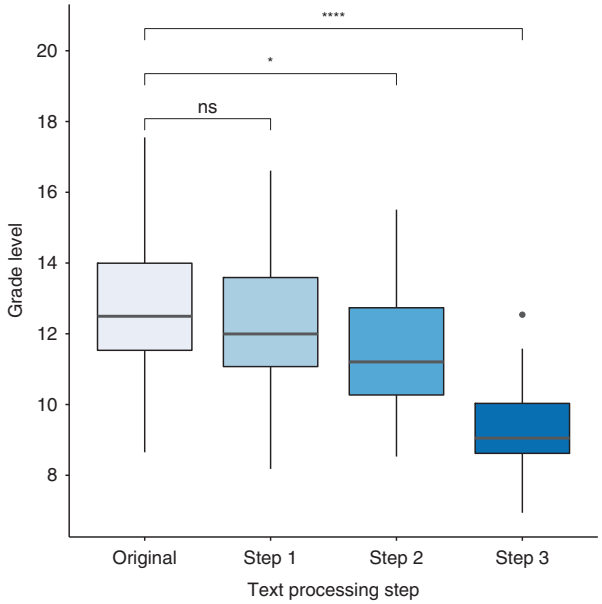

Fig. 4 Reading grade levels of text on American College of Medical Genetics and Genomics (ACMG™) conditions ($n = 28$) after text processing techniques were applied. Original: original Genetics Home Reference (GHR) texts. Step 1: removed medical jargon. Step 2: replaced repeated condition name. Step 3: automatically generate text with key data extracted from GHR, MedlinePlus, ClinGen, and other sources. Kruskal–Wallis test was performed to test the significance of the differences in readability after each process (*ns* not significant; $*p < 0.05$; $****p < 0.0001$).

Our study found that genetic educational materials are complex and are often available in a form that is difficult to read for the general public. When we reviewed commonly used online resources for a genetic condition, most sources had a reading grade level beyond what is recommended. Even though some of these resources are designed to be consumer-targeted, most of the entries had reading levels that far exceed the level that many consumers can understand. This implies that a significant portion of consumers may still be unable to fully utilize consumer health resources to gather information and make informed personal decisions. Patients with limited health literacy, up to 36% of adults in the United States, are especially vulnerable to poorly informed choices, anxiety, and suboptimal medical treatment.[26] Our study supports the need to create online health resources that are more inclusive of diverse literacy levels of consumers and associated socioeconomic backgrounds.

It can be challenging to balance scientific details with simplified information for the public, particularly in a technical field with tremendous depth and detail.[27] Keeping up with the changes in knowledgebase is another challenge. Online health information has made updates easier; however, the volume and complexity of information is substantial. For genetic information to be meaningful and useful to a broader population, it is necessary to provide baseline-level information that ensures wide understanding. For more advanced consumers, technical resources can be readily provided on top of the baseline information.

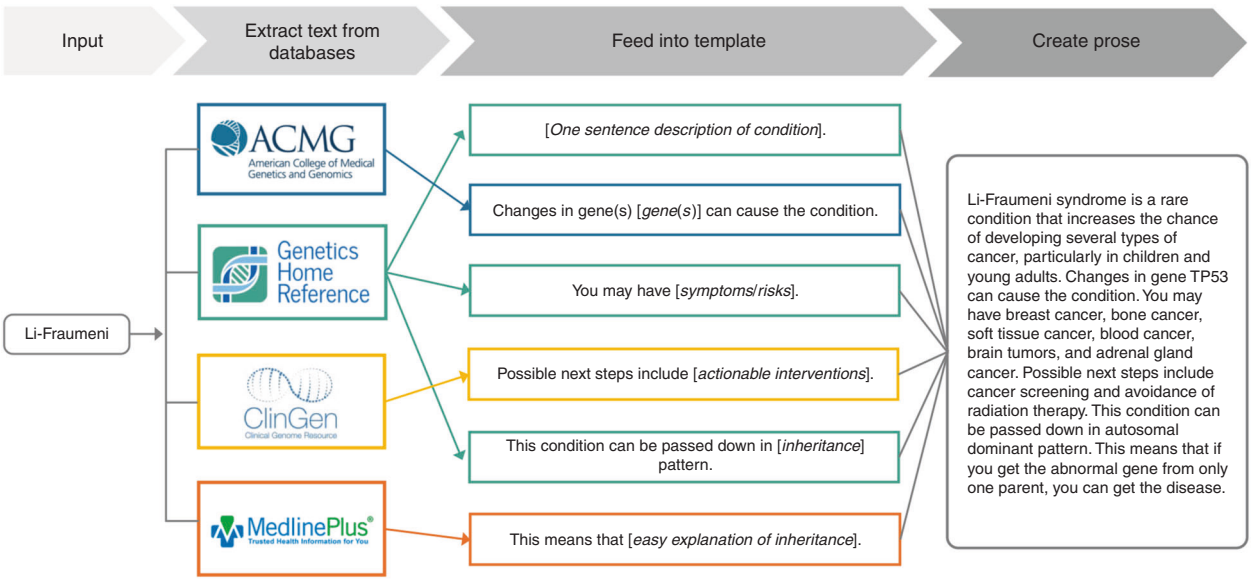

Fig. 5 Flowchart of automated text generation on Li–Fraumeni syndrome using American College of Medical Genetics and Genomics (ACMG™) and National Institutes of Health (NIH) resources.

We found that factors that increase reading difficulty are long sentences, difficult words, and medical jargon. These components can be cut significantly without losing emphasis on patient actionability. We posit that prioritizing medical details in a structured fashion, while using short sentences and simple vocabulary, is a way to reduce complexity. Details can be prioritized with information that patients are most interested in, such as management and next steps regarding their conditions.[28] Different forms of media such as videos and illustrations can be also utilized to make information easier to understand for diverse consumers. We propose that text readability could be improved in a scalable, automated fashion using NLP tools and public databases. This process can be utilized by patient support groups when creating accessible content for diverse readers. In this study, we focused on the actionable reportable conditions list, but these methods could be applied to other genetic conditions, particularly since the list of actionable gene-associated conditions is anticipated to grow as more treatment and management options emerge.

As genomic medicine becomes integrated across medical disciplines in coming years, consumers will increasingly need to access understandable genetic information. Simplifying and providing appropriate genetic health resources will benefit consumers from diverse literacy backgrounds and promote inclusivity. If we can achieve a patient-centered approach that focuses on the individual's context and needs, we can truly achieve success in the personalized genomic era.

SUPPLEMENTARY INFORMATION

The online version of this article (https://doi.org/10.1038/s41436-019-0695-7) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

This research was supported in part by NIH/National Human Genome Research Institute (NHGRI) 5U01HG009599. ACMG™, ACMG SF™, ACMG 59™, ACMG 56™, and related words and designs incorporating ACMG™, are trademarks of American College of Medical Genetics and Genomics and may not be used without permission.

DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. World Wide Web: Pew Research Center. Internet/broadband fact sheet. 2019. https://www.pewinternet.org/fact-sheet/internet-broadband/. Accessed 1 Oct. 2019.
2. Hibbard JH. Patient activation and the use of information to support informed health decisions. Patient Educ Couns. 2017;100:5–7.
3. Lee KC, Berg ET, Jazayeri HE, et al. Online patient education materials for orthognathic surgery fail to meet readability and quality standards. J Oral Maxillofac Surg. 2019;77:180.e1–180.e8.
4. Santos PF, Daar DA, Badeau A, et al. Readability of online materials for Dupuytren's contracture. J Hand Ther. 2017;31:1–7.
5. Sheppard ED, Hyde Z, Florence MN. et al. Improving the readability of online foot and ankle patient education materials. Foot Ankle Int. 2014;35:1282–1286.
6. World Wide Web: Agency for Healthcare Research and Quality, AHRQ Health Literacy Universal Precautions Toolkit, 2nd Edition, 2019. https://www.ahrq.gov/health-literacy/quality-resources/tools/literacy-toolkit/index.html. Accessed 27 Sep. 2019.
7. World Wide Web: National Institutes of Health, Clear Communication: Clear & Simple: 2018. https://www.nih.gov/institutes-nih/nih-officedirector/office-communications-public-liaison/clear-communication/clear-simple. Accessed 5 Sep. 2019.
8. Berkman ND, Sheridan SL, Donahue KE, et al. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med. 2011;155:97–107.
9. Emery J, Kumar S, Smith H. Patient understanding of genetic principles and their expectations of genetic services within the NHS: a qualitative study. Community Genet. 1998;1:78–83.
10. Taylor MRG, Alman A, Manchester DK. Use of the Internet by patients and their families to obtain genetics-related information. Mayo Clin Proc. 2001;76:772–776.
11. Bernstam EV, Shelton DM, Walji M. Instruments to assess the quality of health information on the World Wide Web: what can our patients actually use?. Int J Med Inform. 2005;74:13–19.
12. Mitchell JA, Mccray AT. The Genetics Home Reference: A New NLM Consumer Health Resource. AMIA Annu Symp Proc. 2003:936.
13. US National Library of Medicine. MedlinePlus. 2019. https://medlineplus.gov/.
14. Lewis J, Snyder M, Hyatt-Knorr H. Marking 15 years of the Genetic and Rare Diseases Information Center. Transl Sci Rare Dis. 2017;2:77–88.
15. Pagon RA. GeneTests: an online genetic information resource for health care providers. J Med Libr Assoc. 2006;94:343–348.
16. World Wide Web: March of Dimes. PKU (phenylketonuria) in your baby. 2013. https://www.marchofdimes.org/complications/phenylketonuria-inyour-baby.aspx. Accessed 1 July 2019.
17. World Wide Web: PKU.com. With PKU, the foods you eat directly impact the way your brain functions. https://www.pku.com/about-pku/phein-the-brain. Accessed 1 July 2019.
18. World Wide Web: Mayoclinic.org. Phenylketonuria (PKU). Mayo Clinic. Phenylketonuria (PKU). 2018. https://www.mayoclinic.org/diseases-conditions/phenylketonuria/symptoms-causes/syc-20376302. Accessed 1 July 2019.
19. World Wide Web: National PKU Alliance. About PKU. https://www.npkua.org/What-is-PKU/About-PKU. Accessed 1 July 2019.
20. World Wide Web: National PKU News. What is PKU? 2019. https://pkunews.org/what-is-pku/. Accessed 1 July 2019.
21. World Wide Web: Li–Fraumeni Syndrome Association. 2019. https://www.lfsassociation.org/. Accessed 1 July 2019.
22. World Wide Web: Living LFS. 2018. https://www.livinglfs.org/. Acceessed 1 July 2019.
23. World Wide Web: Cancer.net information from American Society of Clinical Oncology. 2019. http://www.cancer.net/cancer-types/li-fraumeni-syndrome. Acceessed 17 Sep 2019..
24. Gonzalez-Hernandez G, Sarker A, O'Connor K. et al. Advances in text mining and visualization for precision medicine. Biocomputing. 2018;23:559–565.
25. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMGTM SFv2.0): A policy statement of the American College of Medical Genetics and Genomics. Genet Med. 2017;19:249–255.
26. Mills R, Powell J, Barry W. et al. Information-seeking and sharing behavior following genomic testing for diabetes risk. J Genet Couns. 2015;24:58–66.
27. Mitchell J, Fun J, McCary A. Design of Genetics Home Reference: a new NLM consumer health resource. J Am Med Informatics Assoc. 2004;11:439–447.
28. Walser S, Werner-Lin A, Mueller R, et al. How do providers discuss the results of pediatric exome sequencing with families? Per Med. 2017;14:409–422.