

# How far has Optical Character Recognition Softwares have come.\*

## Evaulation of the Accuracy of OCR

Sampson Zhao

30 April 2022

### Abstract

Though the earlier versions of OCR required the text to have a set font for recognition, modern iterations of this has developed enough to recognize multiple fonts as well as different writing systems. This project focuses on the accuracy of the modern OCR system and will compare the accuracy of the produced text to the originals. It was found that without any level of image processing, the OCR would typically catch a lot of blemishes on the paper, resulting in illegible text. While using Magick and a third party image processing software, there was a huge increase in accuracy, with texts very near the 100% accuracy rate. While OCR is still immature, as it is unable to recognize the differences between some letters and blemishes, it is still a work in progress, as developement of A.I. and other text recognition improvements will allow the software to evolve once again.

**Keywords:** OCR, Tesseract, Magick, Accuracy, Historical Texts, Improvements in OCR, AI related to OCR.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Data</b>	<b>2</b>
2.1	Acknowledgement . . . . .	2
2.2	Where this data comes from . . . . .	3
2.3	Data Selection and Concern . . . . .	3
2.4	Understanding the Data: How accurate is OCR? . . . . .	5
<b>3</b>	<b>Discussion</b>	<b>7</b>
3.1	Problems with OCR . . . . .	7
3.2	Progress with OCR . . . . .	9
3.3	Further Development of OCR alongside the development of A.I. . . . .	9
<b>4</b>	<b>Conclusion</b>	<b>9</b>
<b>5</b>	<b>Appendix</b>	<b>10</b>
<b>References</b>		<b>14</b>

---

\*Code and data are available at: <https://github.com/Sampsonzhao/OCR-final>.

# 1 Introduction

OCR was first developed as a technology in the 1910s, as it was primarily used in tandem with reading devices for the blind “The History of OCR, Optical Character Recognition : Schantz, Herbert f : Free Download, Borrow, and Streaming” (1982), with the main premise of OCR being using some optical device to capture characters on a document, and transform it into a manner where computer software can then recognize the text. Though earlier iterations of OCR were used in the development of text documents to audio readers for the blind or visually impaired, the technology is now widely used as a shortcut in data entry, as documents like passports, cheques, and other printed documentations, can be quickly and accurately transformed, effectively digitizing the document to allow for editing and searching electronically. Though earlier iterations of the software required documents to be in particular fonts for the software to recognize the text, more modern and advanced systems have developed significantly, to the point where these software are able to recognize texts in a variety of fonts within the same document, as well as recognizing different writing systems, like Cyrillic, Arabic, etc. (List of most of the available languages that are recognized by Google’s OCR software here).

OCR development has been progressing fairly well, as the accuracy of the software from 19th century to the 20th century has seen an increase of 81% accuracy to 99% accuracy. Though this is comparing the developments of the technologies that these softwares are based off of, there was a fundamental shift in the style of recognition. OCR developed from a character-to-character recognition style to a more pattern recognition style, signifying the advancements of computational power, as effectively they changed from a character-to-character recognition to more of a line-by-line recognition. Though the recognition of text improved drastically back then, there were multiple variables that caused the 1% of texts to fail, and that is related to the pre-processing of the image. Factors like wrinkles, slanting, normalization of ratio and scales, were all factors that attributed to failed OCR reads, these were considered artifacts that the software could not recognize, effectively showing garbage code as a result. Though these factors have improved since the early 2000s, OCR is still continuing on the development of optimizing the overall recognition software.

Ultimately, this paper looks into the development of OCR software that are open access today, effectively looking at the accuracy of the text that is produced by the OCR software, when compared to the original text. We are able to use this information to track the accuracy of these software, which would be useful in digitizing older texts. Here, we would look at the difference in processed and unprocessed images of text documents from different titles and compare the accuracy of OCR, as well as testing the limits of OCR. The continued development of OCR aims to do this as one of its main applications, as it would allow for further digitization of older texts, allowing more of this information to be shared on the web. As many older texts have only photo renditions of the book online, OCR can be used to scan these pictures and create a searchable and interactive version of the text.

While many old texts show faded or blurred text, due to the age of the documents themselves, this posed an issue for both old and newer versions of OCR, as the pre-processing of the image would result in the software returning errors alongside the text. There are tweaks that are currently in development, as development of application-specific OCR has come up as one of the main projects developed by government entities, like recognizing license plate numbers or invoice specifics. These developments allow for the specialization of the software, which trims down on the processing time needed for the software to recognize these texts. The hope is that further development of application-specific OCR and effectively transition to further optimizing the recognition of older texts, allowing for the continued digitization of historical texts.

# 2 Data

## 2.1 Acknowledgement

This investigation uses the ‘R’ programming language (R Core Team 2020), and the ‘R-Studio’ Integrated development environment (RStudio Team 2020) as the primary source of data processing and the generation of the paper. The ‘Tidyverse’ package was used for manipulation of the datasets (Wickham et al. 2019). The

‘Here’ package was used to locate files produced from R scripts (Müller 2020). ‘KableExtra’ package was used to enhance the overall appearance of tables(Zhu 2021). ‘knitr’ was used to tie all of the information on this document into one digestible PDF [Xie (2021)][Xie 2015](Xie 2014). ‘ggplot2’ was used to create the bar plots in this document (Wickham 2016). ‘reshape2’ was used to modify data sets to better create the wanted plots, while simplifying the code needed for those graphs (Wickham 2007). ‘Tesseract’, ‘magick’, and ‘magrittr’ was used to implement the OCR recognition of the images, as well as slight modifications of the image for better recognition[Ooms (2022)][Ooms 2021](Bache and Wickham 2020). ‘diffr’ was used to generate plots that compared the text files produced by the OCR from the previously mentioned packages and provide a visual comparison of two text documents(Muschelli 2017). ‘RecordLinkage’ and ‘fuzzywuzzyR’ were both used to generate a percentage difference between two text documents using different algorithms[Sariyar and Borg (2022)][(Mouselimis 2021)(Inc 2014)]. ‘reticulate’ was used to install the correct python environment for package ‘fuzzywuzzyR’ to function properly(Ushey, Allaire, and Tang 2022).

## 2.2 Where this data comes from

While there is no direct rational behind the use of these particular books for this investigation, the overall goal of this investigation was to look at texts that were on different parts of the spectrum. Two books were used to represent the opposite sides of the spectrum of time, as well as a standard. In this investigation, the following books were used as sources for samples for the OCR:

- Book 1: “Basic and Clinical Pharmacology 14th Edition” by Bertram E. Katzung
- Book 2: “Life and Literature Book One Grade 7” by The Ministry of Education For Ontario

Page 1 of the following book was used as a standard: \* “The Life and Work of Fredson Bowers” by G. Thomas Tanselle

“Basic and Clinical Pharmacology 14th Edition” (Katzung and Katzung 2018) represents more recent texts that have clean and non-intrusive text, alongside the fact that a digital copy of this book is available to retrieve the original text for comparison. “The Life and Work of Fredson Bowers” (Tanselle and Battestin 1993) was chosen as the standard, as it was commonly used as a standard for testing OCR, as it was seen to be used in multiple sources that talked about the uses of OCR and tesseract [(2022)][(Ooms 2017)]. “Life and Literature” then represents the other end of the spectrum that was described in the introduction, looking at texts that may have faded texts due to aging as well as wear-and-tear, causing the text to become illegible in some cases. Overall, the goal was met, where we have samples from different parts of texts that are still in circulation in some public libraries.

## 2.3 Data Selection and Concern

In this report, due to the sheer size of the books that were chosen as the samples for OCR (The Pharmacology Textbook has over 1200 pages in it, where as Life and Literature has over 300), 5 pages from each book were selected to represent a portion of the book. To remove any related bias, a random number generator was used to select page numbers for each book (except the standard, as that uses page 1 of the book.) The use of a random number generator makes sure that the page numbers selected in succession have no common variable between them, removing any influence from the writer, and thus removing a source of bias (Haahr 2019). For the case of the Pharmacology Textbook, because the pages have text split in half and to decrease the amount of artifacts that the OCR scans for, half of the page would also be selected. This would be done in the same manner as the random number generator used for the page selection of the books, where 1 would represent the left half of the page, and 2 would represent the right half of the page. Hence for this investigation, the following were chosen for OCR processing:

- Book 1: “Basic and Clinical Pharmacology 14th Edition” by Bertram E. Katzung
  - Page 100, left half of page

- Page 259, left half of page
- Page 493, right half of page
- Page 613, right half of page
- Page 1099, right half of page
- Book 2: “Life and Literature Book One Grade 7” by The Ministry of Education For Ontario
  - Page 65
  - Page 127
  - Page 183
  - Page 235
  - Page 318

For the rest of the paper, the books will be referred to as ‘Book 1’ and ‘Book 2’.

For this investigation, the selected pages have been digitized via photo taken from an iPhone X. (These photos can be found in the /input/data folder of the github repository. These photos would then be cropped to limit the amount of white spaces and artifacts that would be introduced into the produced OCR text file. Three levels of processing will be tested:

- Raw/Unprocessed
  - This level would have no digital processing applied to the photo before running the photo through OCR. Cropping will be done on these photos, as it was found that if no cropping was done, the percentage of recognizable text would be very near 0%.
- Some Processing/Cleaning
  - This level will have some digital processing from the built-in image processing methods of Magick. This will be limited to converting the photo to gray-scale to limit the amount of artifacts picked up by OCR, enlarging the photo, increased brightness and contrast to attempt to remove shadows, as well as trimming additional white space.
- Advanced Processing/Cleaning
  - This level contains a larger amount of digital processing, as the photos were put through a 3rd-party software (The software used in this investigation was ‘Scanner Pro’ by Readdle Technologies Limited). The software is able to produce scanner-like images based off of the photos that are uploaded to the software, and is able to correct warping of the page, get rid of any shadows that are in the raw photo, as well as boosting the brightness and contrast to ensure that almost all text is clear and visible. While this is no different from physically scanning the pages through a printer/scanner, these were not available at the time of this investigation due to a malfunction.

These photos will then undergo a script, which will take the photo and process them based on their respective level. In the case of the photos that underwent the advanced image processing level, those will be processed on the 3rd-party app, then imported into the /inputs/data/ folders, as this app does not have an R related package that can be used. After the processing of the photos, they will run through the built-in OCR package in Magick, and the collected text will be exported, and stored for further analysis. This exported text data would then be compared to the original text, that is either collected through a digital copy of the book, or typed out manually. The comparison will produce both a visual comparison as well as a percent difference to the original, which would then be exported and stored. All of the previously mentioned files and scripts are available [here](#).

The use of the standard was to confirm that the OCR software would be working properly, as it was used to confirm that the accuracy percentage of a scanned document between raw, semi-processed and fully processed photos would be similar, as the documentation would require no additional processing, and applying these modifications to the original photo would be negligible, in terms of the change in accuracy. “The Life and Work of Fredson Bowers” by G. Thomas Tanselle would act as a positive control, while a blank piece of paper would act as a negative control.

## 2.4 Understanding the Data: How accurate is OCR?

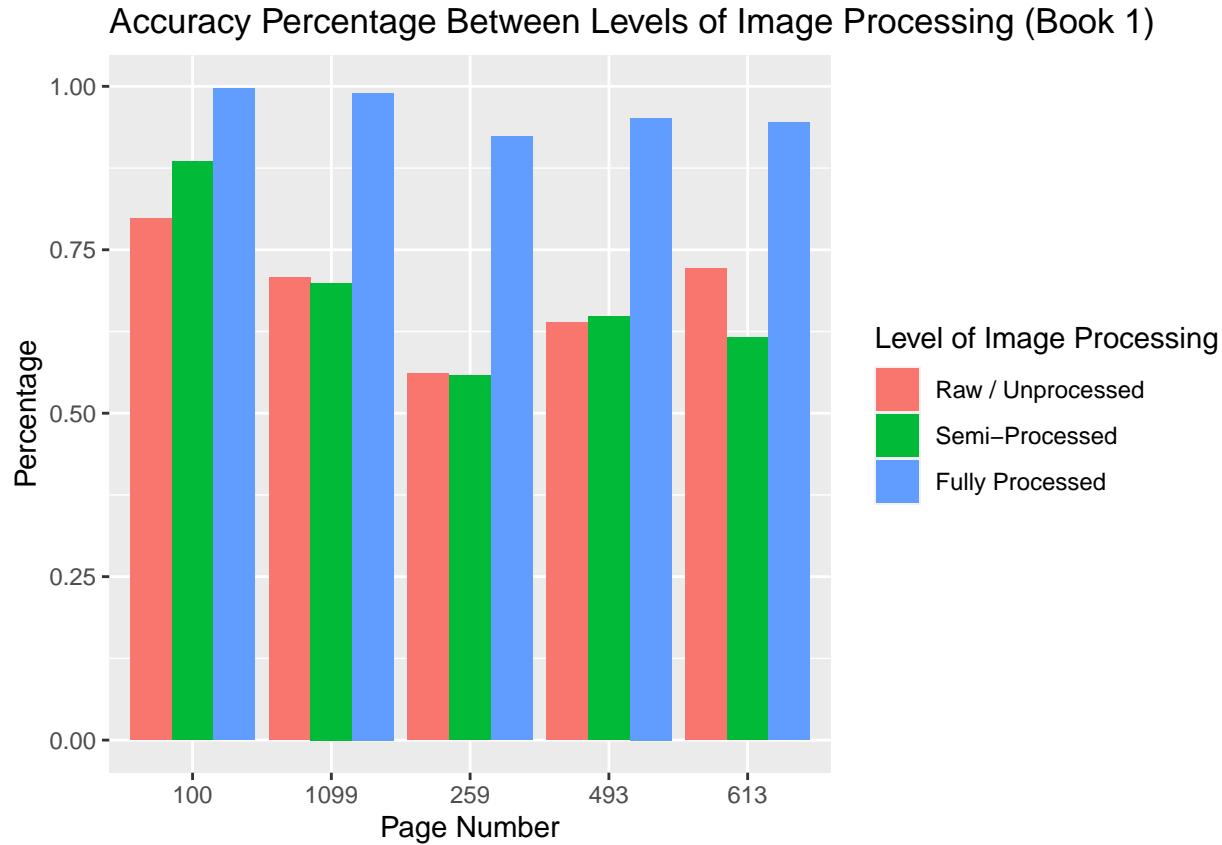


Figure 1: Change in Accuracy Based on image processing For Book 1

In Figure 1, the investigation reveals a similar trend that is common in most of the samples that were tested. There is a positive correlation between the amount of processing done on the raw photo and the accuracy of the exported text after running the photo through OCR. While this is to be expected, we can also see that this was not strictly followed, as there were some that saw no change after a small amount of image processing, most notable for Figure 1 would be Page 1099 and page 259, as these two pages showed no increase in accuracy after processing. Page 613 showed an extreme reverse in this case, as after processing the photo through Magick, it was found to have a lower accuracy, compared to the raw photo. This phenomenon was seen multiple times through both Book 1 and Book 2. A closer investigation was done on Page 613 (As seen in 4 and 5), and it shows that there are errors throughout the entire text document of both the semi-processed and raw photos. Based on these comparisons, it can be assumed that the modification that were done on the raw photo through Magick may have created more artifacts, which in turn allowed the OCR software to pick up more errors, causing the decrease in accuracy after processing. This event may look like an outlier case, but there may be other factors that affect the accuracy of the OCR, which discussed in Section 3.1.

## Accuracy Percentage Between Levels of Image Processing (Book 2)

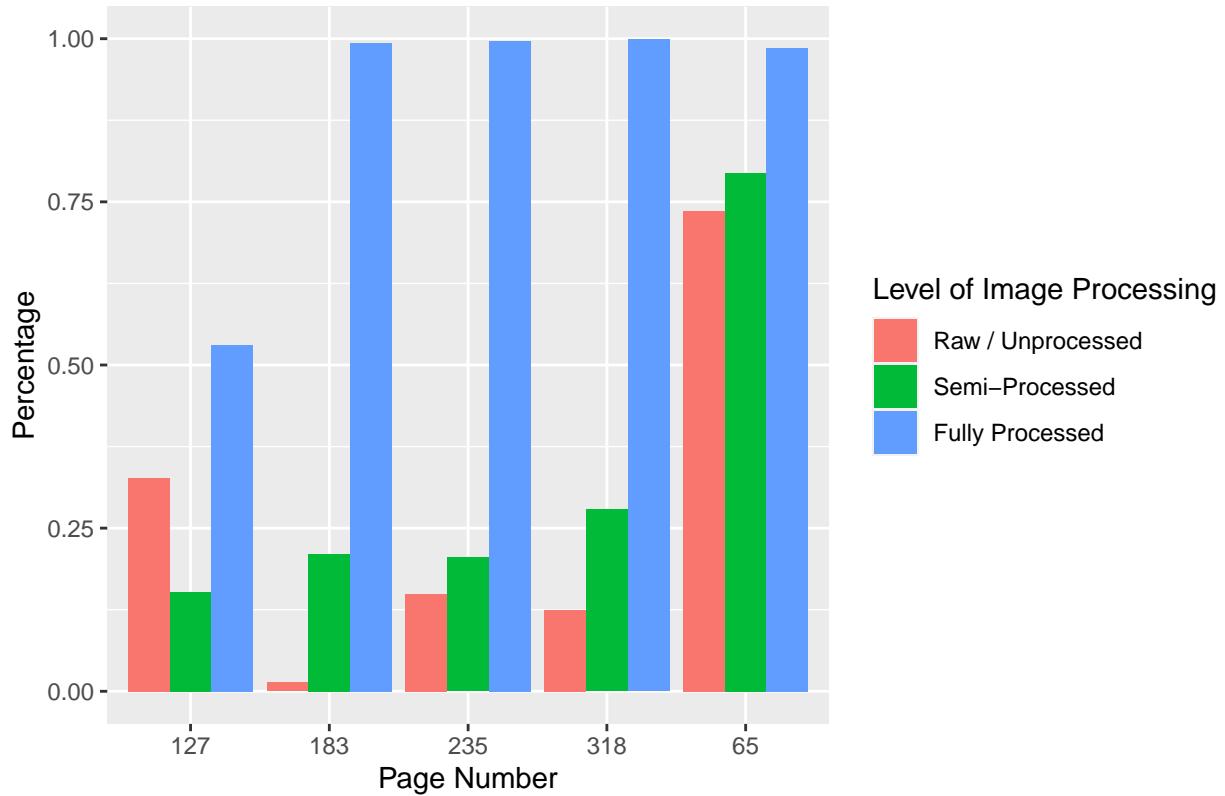


Figure 2: Change in Accuracy Based on image processing For Book 2

In the case of Figure 2, we can see that there are examples of more extreme cases of the positive correlation between image processing and OCR accuracy, as most of the samples in book 2 show very little text being extracted from the photo without any processing. After some processing of the photo, we could see a fair increase in accuracy, when compared to the raw photo, and finally, after removing many of the artifacts through the advanced image processing software, we can see that most documents have a near 100% accuracy rate. While there is still an outlier, in sample Page 127, where we see the decrease in accuracy after some level of processing, a similar conclusion can be said about this outlier like the previous outlier, mentioned in the Figure 1 description. The changes done through Magick may have faded some of the text or warped the text, leading to the OCR scan not being able to recognize the photo properly, causing the decrease in accuracy. The main difference that we can see between Figure 1 and figure 2 is the relatively low accuracy rate throughout all of the raw/unprocessed samples. This distinction is relative to the age of the original texts themselves, as different physical characteristics of book 2 can attribute to the lower accuracy rate. If we compare the raw photo for Book 1 and Book 2 (Shown in Appendix Figure 6 and 7), we can see that Book 2 shows a yellowish paper, with variations in the texture of the paper, while book 1 shows a much more uniform texture. While these factors may not be significant to evaluating the accuracy of OCR, it shows some of the limitations of OCR when it comes to dealing with older texts.

Page	Raw	Basic Processing	Advanced Processing	Page	Raw	Basic Processing	Advanced Processing
100	79.8%	88.5%	99.73%	65	73.6%	79.42%	98.54%
259	56.0%	55.8%	92.30%	127	32.7%	15.22%	53.06%
493	63.8%	64.7%	95.17%	183	1.4%	21.07%	99.31%
613	72.1%	61.6%	94.46%	235	14.9%	20.54%	99.65%
1099	70.8%	69.9%	98.98%	318	12.5%	27.85%	99.90%
Average	68.5%	68.1%	96.13%	Average	27.0%	32.82%	90.09%

Table 1: Percent Representation of OCR scans for Book 1 (Left) and Book 2 (Right).

Table 1 above represents the numerical representation of the graphs seen in figure 1 and figure 2. While the information that is shown on this table is the same as the aforementioned graphs, this table also gives a relative view of comparing the average accuracy for each book. We can see that while both books eventually reach a very high level of accuracy (over 90% for both books) after enough image processing, we can see the relative ineffectiveness of the basic processing, when compared to the raw photo. In the case of both books, we see that the basic image processing provided by Magick is unable to increase the accuracy of the OCR scan, and in book 1, it seemed to have a negative effect on accuracy as well. While this may be attributed to the photo being over-trimmed, leading issues with spacing between each letter, we can also look into comparing the averages between the two books. We can see that most likely due to the age of the book and an increase in artifacts and marks on the pages of book 2, we see that the average accuracy rates for all three levels of processing are considerably lower than that of book 1. This may be due to a number of different factors, but most notably, the difference of paper type (book 1 tend to have more glossy paper, while book 2 had more textured/matte paper) may have caused some issues relating to how OCR recognizes blemishes, leading to the decrease in accuracy.

## 3 Discussion

### 3.1 Problems with OCR

While OCR is a very convenient piece of software that allows us to quickly translate information from a photo into an interactable piece of digital text, there are still many flaws associated to the software itself. For example, while dealing with the images for Book 1, there was a issue related to the overall format of the text. OCR is not able to differentiate between full-page texts and split page text, hence it does not recognize the difference between them. This would then cause errors for the right half of the text, as the spacing between the two bodies of text throws off the OCR recognition, causing the corruption of the data on the right hand side. From this investigation, it was observed that the spacing between each character, as well as the spacing between bodies of text affect how Tesseract views the text itself. While this wouldn't be a big problem in normal circumstances, it would be an issue while looking into the further goals of OCR. The lack of flexibility in OCR, as well as the lack of understanding in sentence structures are one of the downfalls of the software. Though these are easily ameliorated in the form of restricting recognition of text, as well as restricting white spaces, this may become an issue for how OCR can be used in text recovery and text recognition in older texts.

Based on a similar issue as mentioned above, we can also look into the artifacts that are introduced due to how the original text is captured. Though in Section 2.2, it was mentioned that all the photos have some level of cropping done to it. For Book 1 (Basic and Clinical Pharmacology), texts were limited to only half of the page, as there was no continuity in the sentence, as well as OCR not recognizing the white space used to separate the two columns of text. Though that played a large factor in generating large amounts of garbled data in the exported text file, artifacts on the edge of the photo, like the area marked in red in Figure 3, have been found to also cause some errors, as any sort of lines would be recognized as text by OCR. Hence,

in the case of all of the photos that were taken in a similar manner, all of the edges were removed to make sure that there would not be false hits from the OCR scan.

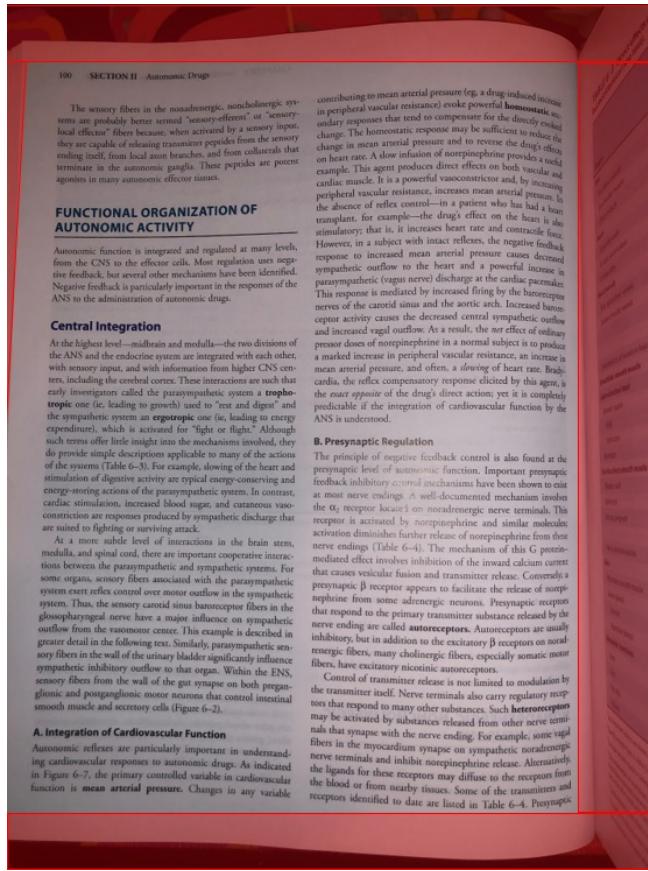


Figure 3: Edge of photo is recognized by Tesseract as text

In addition to cropping photos, there were various image manipulations done to further increase the accuracy of the OCR scans. If we look back at Table 1, we can see how important being able to clean and remove any artifacts from the photo, as we tend to see an positive correlation between cleaning and percent recognition. This brings into light the different photo manipulations needed for a clean scan, and how that might take away from the original text. In the case of comparing the raw photo to the semi-processed photo, we can see that even just removing colour and increasing the brightness and contrast a bit, would generally be enough for the OCR software to recognize the text better. To be more accurate, the hypothesis for this phenomenon is more related to how the OCR software is unable to see flaws and artifacts if the brightness and contrast is increased. By modifying the photo slightly, the software is able to discern better the difference between text and random debris, allowing for increased accuracy in the text output. This is significant, due to the fact that not every photo is as clean as a scanner, as photos generally have large amounts of scrap data in the periphery, causing the OCR to incorrectly identify the characters, sometimes causing OCR scans to show up as errors. The same can be said when looking at the accuracy when more image processing is done on the samples. By completely removing all colour from the images, as well as being able to correct the warping of the image due to the natural curve of the book, the OCR scan was able, in most cases, get a clean scan of the entire document. Based on the accuracy reports in Figure 1 and 2, we can see how significant those changes can be, which supports the implementation of full image processing along side OCR for cleaner and more accurate text extraction.

### **3.2 Progress with OCR**

In general, if this experiment was repeated, a suggestion can be made to increase the accuracy overall, by scanning the pages of texts, as well as preventing large shadows from being shown, as both of these factors have affected the overall percentage difference between the 3 levels of image processing. It is understandable that scanning certain books may not be possible, as it may incur damage to the physical item itself. As the premise of this paper is looking at how OCR can be used in the digitization of older materials. For example Book 2 (Life and Literature), this is an item from 1943. The age of the item as well as the physical condition of the item was considered to be fairly poor and would not have fared well if the spine was overextended. Scanning of many older literature would not be feasible, hence photos would be the next best alternative. While third-party software were available for this investigation, there have been progress in the overall OCR recognition, as there were mentions of OCR being able to discern text from uneven surfaces. While it was not apparent during this investigation, it is something that is progressing in the field of OCR, and would become a very powerful tool for those who wish to continue the digitization of many older texts. If used properly, image processing software could have OCR built into the overall application, allowing for less processing on the OCR side.

Overall, OCR has become a tool that is important in both the professional environment and in every day use, as it is a tool that allows us to quickly scan and produce digital copies of text. People may use this tool to quickly scan and create modifiable versions of these texts, allowing for a much more flexible experience to the material. And in the professional environment, written documents can be scanned and deciphered to produce digital copies of these texts, if the OCR was fed hand-written fonts.

### **3.3 Further Development of OCR alongside the development of A.I.**

There was discussions going on talking about the use of A.I. in relation to OCR, as it would be something that can be used in tandem with OCR to improve the overall accuracy of the software itself. The hypothesis of this theory is that OCR would be able to use A.I. neural learning to collect different variations of fonts, texts, sentence structures, formats, etc. and use the collected information to better build the optical recognition of different fonts. This may also be something that increase accuracy in older texts, A.I. can be implemented in a way to suggest letters that may have been missing due to wear-and-tear, as well as the imperfections from the capture method of the photo. These factors were main concerns when working with the samples in this investigation. Though A.I. has not developed far enough for this use-case, it is something that can help progress the use of OCR in future projects, allowing for clearer and more precise scans, and producing more accurate text representations.

## **4 Conclusion**

OCR is a fascinating piece of software that has existed since the 1910s, starting off as an auxiliary tool for reading devices for the blind, and now, its commonly used in tandem with many pieces of software to produce functionality in many applications. As a tool that has been evolving over the past 110 years, it has allowed people to accurately transform and digitize documents, and allow for them to be edited and searched electronically. While the progression in the development of OCR has slowed down a lot in recent years, the technology is still slowly growing, with the addition of A.I. learning, leading to the increase in accuracy of text outputs. This investigation was able to show that with enough image processing, such as reducing the amount of shadows, increasing the brightness and contrast, as well as removing saturation of the photo, would allow for a near 100% accuracy rate. But this investigation also shows the downfalls of OCR as of right now, as it is not able to discern between a ripple in the page and a letter in the alphabet. OCR itself is still growing to be able to adapt to these artifacts, but for now, it is still a technology that is slowly becoming something that can bring light to books of the past.

## 5 Appendix

Raw / Unprocessed	Original
<pre> 1 CHAPTER 34 Drugs Used in Disorders of Coagulation 613 2 3 Administration &amp; Dosage 4 The indications for the use of heparin are described in the   section 5 on clinical pharmacology. A plasma concentration of heparin   of 6 0.2-0.4 unit/mL (by protamine titration) or 0.3&amp;#033;0.7   unit/mL 7 (anti-Xa units) is considered to be the therapeutic range for   treatment of venous thromboembolic disease. This concentra- 8 tion generally corresponds to a PTT of 1.5-2.5 times   baseline. 9 However, the use of the PTT for heparin monitoring is   problem- 10 atic. There is no standardization scheme for the PTT as there   is 11 for the prothrombin time (PT) and its international   normalized 12 ratio (INR) in warfarin monitoring. The PTT in seconds for a 13 given heparin concentration varies between different reagent/ 14 instrument systems. Thus, if the PTT is used for monitoring,   the 15 laboratory should determine the clotting time that   corresponds to 16 the therapeutic range by protamine titration or anti-Xa   activity, 17 as listed above. 18 In addition, some patients have a prolonged baseline PTT due   to 19 to factor deficiency or inhibitors (which could increase   bleeding 20 risk) or lupus anticoagulant (which is not associated with   bleeding 21 risk but may be associated with thrombosis risk). Using the   PTT 22 to assess heparin effect in such patients is problematic. An   alter- 23 native j &amp;#033;to use anti-Xa activity to assess heparin   concentration, a 24 25 e test now widely available on automated coagulation   instruments. 26 This approach measures   heparin concentration; however, it   does 27 not provide the global assessment of intrinsic pathway   integrity. 28 Continuous intravenous administration of heparin is accom-   plished via an infusion pump. After an initial bolus 29 injection of 30 80&amp;#033;10 units/kg, a continuous infusion of about 15&amp;#033;22   units/kg 31 per hour is required to maintain the anti-Xa activity in the   range of 0.3&amp;#033;0.7 units/mL. Low-dose prophylaxis is achieved 32 with subcutaneous administration of heparin, 5000 units   every 8&amp;#033;12 hours. Because of the danger of hematoma forma- 33 tion at the injection site, heparin must never be   administered 34 intramuscularly. 35 Prophylactic enoxaparin is given subcutaneously in a dosage   of 30 mg twice daily or 40 mg once daily. Full-dose 36 enoxaparin 37 therapy is 1 mg/kg subcutaneously every 12 hours. This cor-   responds to a therapeutic anti-factor Xa level of 0.5&amp;#033;1   unit/mL. 38 Selected patients may be treated with enoxaparin 1.5 mg/kg   once a day, with a target anti-Xa level of 1.5 units/mL. 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 </pre>	<pre> CHAPTER 34 Drugs Used in Disorders of Coagulation 613 1 2 3 Administration &amp; Dosage 4 The indications for the use of heparin are described in the   section 5 on clinical pharmacology. A plasma concentration of heparin   of 6 0.2&amp;#033;0.4 unit/mL (by protamine titration) or 0.3&amp;#033;0.7   unit/mL 7 (anti-Xa units) is considered to be the therapeutic range for   treatment of venous thromboembolic disease. This concentra- 8 tion generally corresponds to a PTT of 1.5&amp;#033;2.5 times   baseline. 9 However, the use of the PTT for heparin monitoring is   problem- 10 atic. There is no standardization scheme for the PTT as there   is 11 for the prothrombin time (PT) and its international   normalized 12 ratio (INR) in warfarin monitoring. The PTT in seconds for a 13 given heparin concentration varies between different reagent/ 14 instrument systems. Thus, if the PTT is used for monitoring,   the 15 laboratory should determine the clotting time that   corresponds to 16 the therapeutic range by protamine titration or anti-Xa   activity, 17 as listed above. 18 In addition, some patients have a prolonged baseline PTT due   to 19 to factor deficiency or inhibitors (which could increase   bleeding 20 risk) or lupus anticoagulant (which is not associated with   bleeding 21 risk but may be associated with thrombosis risk). Using the   PTT 22 to assess heparin effect in such patients is problematic. An   alter- 23 native is to use anti-Xa activity to assess heparin   concentration, a 24 test now widely available on automated coagulation   instruments. 25 This approach measures heparin concentration; however, it   does 26 not provide the global assessment of intrinsic pathway   integrity   of the PTT. 27 The following strategy is recommended: prior to initiating   therapy, an initial bolus of 28 5000 units of heparin is administered   intravenously over 15 minutes. This dose 29 maintains a therapeutic anti-Xa level of 0.3&amp;#033;0.7 units/mL   for approximately 12 hours. Subsequent doses of 30 5000 units are administered   every 12 hours. Because of the danger of hematoma forma- 31 tion at the injection site, heparin must never be   administered 32 intramuscularly. 33 Prophylactic enoxaparin is given subcutaneously in a dosage   of 30 mg twice daily or 40 mg once daily. Full-dose 34 enoxaparin 35 therapy is 1 mg/kg subcutaneously every 12 hours. This cor-   responds to a therapeutic anti-factor Xa level of 0.5&amp;#033;1   unit/mL. 36 Selected patients may be treated with enoxaparin 1.5 mg/kg   once a day, with a target anti-Xa level of 1.5 units/mL. 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 </pre>

Figure 4: Comparison of Raw Photo OCR to Original of Page 613 of Book 1

Semi-Processed	Original
<pre> 1 CHAPTER 34 Drugs Used in Disorders of Coagulation 613 2 Administration &amp; Dosage 3 The indications for the use of heparin are described in the   section 4 on clinical pharmacology. A plasma concentration of heparin   of 5 0.2±0.4 unit/mL (by protamine titration) or 0.3±0.7   unit/mL 6 (anti-Xa units) is considered to be the therapeutic range for 7 treatment of venous thromboembolic disease. This concentra- 8 tion generally corresponds to a PTT of 1.5±2.5 times 9 baseline. 10 However, the use of the PTT for heparin monitoring is   problem- 11 atic. There is no standardization scheme for the PTT as there   is 12 for the prothrombin time (PT) and its international   normalized 13 ratio (INR) in warfarin monitoring. The PTT in seconds for a 14 given heparin concentration varies between different reagent/ 15 instrument systems. Thus, if the PTT is used for monitoring,   the 16 laboratory should determine the clotting time that   corresponds to 17 the therapeutic range by protamine titration or anti-Xa   activity, 18 as listed above. 19 In addition, some patients have a prolonged baseline PTT due   to factor deficiency or inhibitors (which could increase   bleeding 20 risk) or lupus anticoagulant (which is not associated with   bleeding 21 task but may be associated with thrombosis risk). Using the   PTT 22 to assess heparin effect in such patients is problematic. An   alter- 23 native is to use anti-Xa activity to assess heparin   concentration, a 24 test now widely available on automated coagulation   instruments. 25 This approach measures heparin concentration; however, it does   not provide the global assessment of intrinsic pathway 26 integrity. 27 Continuous intravenous administration of heparin is accom-   plished via an infusion pump. After an initial bolus 28 injection of 29 80±100 units/kg, a continuous infusion of about 15±22   units/kg 30 per hour is required to maintain the anti-Xa activity in the   range of 0.3±0.7 units/mL. Low-dose prophylaxis is achieved 31 with subcutaneous administration of heparin, 5000 units   every 8±12 hours. Because of the danger of hematoma forma- 32 tion at the injection site, heparin must never be   administered 33 intramuscularly. 34 Prophylactic enoxaparin is given subcutaneously in a dosage   of 30 mg twice daily or 40 mg once daily. Full-dose 35 enoxaparin 36 therapy is 1 mg/kg subcutaneously every 12 hours. This cor-   responds to a therapeutic anti-factor Xa level of 0.5±1   unit/mL. 37 Selected patients may be treated with enoxaparin 1.5 mg/kg   once a day, with a target anti-Xa level of 1.5 units/mL. 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 </pre>	<pre> CHAPTER 34 Drugs Used in Disorders of Coagulation 613 Administration &amp; Dosage The indications for the use of heparin are described in the section on clinical pharmacology. A plasma concentration of heparin of 0.2±0.4 unit/mL (by protamine titration) or 0.3±0.7 unit/mL (anti-Xa units) is considered to be the therapeutic range for treatment of venous thromboembolic disease. This concentra- tion generally corresponds to a PTT of 1.5±2.5 times baseline. However, the use of the PTT for heparin monitoring is problem- atic. There is no standardization scheme for the PTT as there is for the prothrombin time (PT) and its international normalized ratio (INR) in warfarin monitoring. The PTT in seconds for a given heparin concentration varies between different reagent/ instrument systems. Thus, if the PTT is used for monitoring, the laboratory should determine the clotting time that corresponds to the therapeutic range by protamine titration or anti-Xa activity, as listed above. In addition, some patients have a prolonged baseline PTT due to factor deficiency or inhibitors (which could increase bleeding risk) or lupus anticoagulant (which is not associated with bleeding risk but may be associated with thrombosis risk). Using the PTT to assess heparin effect in such patients is problematic. An alternative is to use anti-Xa activity to assess heparin concentration, a test now widely available on automated coagulation instruments. This approach measures heparin concentration; however, it does not provide the global assessment of intrinsic pathway integrity. Continuous intravenous administration of heparin is accom- plished via an infusion pump. After an initial bolus injection of 80±100 units/kg, a continuous infusion of about 15±22 units/kg per hour is required to maintain the anti-Xa activity in the range of 0.3±0.7 units/mL. Low-dose prophylaxis is achieved with subcutaneous administration of heparin, 5000 units every 8±12 hours. Because of the danger of hematoma forma- tion at the injection site, heparin must never be administered intramuscularly. Prophylactic enoxaparin is given subcutaneously in a dosage of 30 mg twice daily or 40 mg once daily. Full-dose exoxaparin therapy is 1 mg/kg subcutaneously every 12 hours. This cor- responds to a therapeutic anti-factor Xa level of 0.5±1 unit/mL. Selected patients may be treated with enoxaparin 1.5 mg/kg once a day, with a target anti-Xa level of 1.5 units/mL. </pre>

Figure 5: Comparison of Semi-Processed Photo OCR to Original of Page 613 of Book 1

The sensory fibers in the nonadrenergic, noncholinergic systems are probably better termed "sensory-efferent" or "sensory-local effector" fibers because, when activated by a sensory input, they are capable of releasing transmitter peptides from the sensory ending itself, from local axon branches, and from collaterals that terminate in the autonomic ganglia. These peptides are potent agonists in many autonomic effector tissues.

### FUNCTIONAL ORGANIZATION OF AUTONOMIC ACTIVITY

Autonomic function is integrated and regulated at many levels, from the CNS to the effector cells. Most regulation uses negative feedback, but several other mechanisms have been identified. Negative feedback is particularly important in the responses of the ANS to the administration of autonomic drugs.

#### Central Integration

At the highest level—midbrain and medulla—the two divisions of the ANS and the endocrine system are integrated with each other, with sensory input, and with information from higher CNS centers, including the cerebral cortex. These interactions are such that early investigators called the parasympathetic system a **trophotropic** one (ie, leading to growth) used to "rest and digest" and the sympathetic system an **ergotropic** one (ie, leading to energy expenditure), which is activated for "fight or flight." Although such terms offer little insight into the mechanisms involved, they do provide simple descriptions applicable to many of the actions of the systems (Table 6-3). For example, slowing of the heart and stimulation of digestive activity are typical energy-conserving and energy-storing actions of the parasympathetic system. In contrast, cardiac stimulation, increased blood sugar, and cutaneous vasoconstriction are responses produced by sympathetic discharge that are suited to fighting or surviving attack.

At a more subtle level of interactions in the brain stem, medulla, and spinal cord, there are important cooperative interactions between the parasympathetic and sympathetic systems. For some organs, sensory fibers associated with the parasympathetic system exert reflex control over motor outflow in the sympathetic system. Thus, the sensory carotid sinus baroreceptor fibers in the glossopharyngeal nerve have a major influence on sympathetic outflow from the vasomotor center. This example is described in greater detail in the following text. Similarly, parasympathetic sensory fibers in the wall of the urinary bladder significantly influence sympathetic inhibitory outflow to that organ. Within the ENS, sensory fibers from the wall of the gut synapse on both preganglionic and postganglionic motor neurons that control intestinal smooth muscle and secretory cells (Figure 6-2).

#### A. Integration of Cardiovascular Function

Autonomic reflexes are particularly important in understanding cardiovascular responses to autonomic drugs. As indicated in Figure 6-7, the primary controlled variable in cardiovascular function is **mean arterial pressure**. Changes in any variable

contributing to mean arterial pressure (eg, a drug-induced increase in peripheral vascular resistance) evoke powerful **homeostatic** secondary responses that tend to compensate for the directly evoked change in mean arterial pressure and to reverse the drug's effects on heart rate. A slow infusion of norepinephrine provides a useful example. This agent produces direct effects on both vascular and cardiac muscle. It is a powerful vasoconstrictor and, by increasing peripheral vascular resistance, increases mean arterial pressure. In the absence of reflex control—in a patient who has had a heart transplant, for example—the drug's effect on the heart is also stimulatory; that is, it increases heart rate and contractile force. However, in a subject with intact reflexes, the negative feedback response to increased mean arterial pressure causes decreased sympathetic outflow to the heart and a powerful increase in parasympathetic (vagus nerve) discharge at the cardiac pacemaker. This response is mediated by increased firing by the baroreceptor nerves of the carotid sinus and the aortic arch. Increased baroreceptor activity causes the decreased central sympathetic outflow and increased vagal outflow. As a result, the *net* effect of ordinary pressor doses of norepinephrine in a normal subject is to produce a marked increase in peripheral vascular resistance, an increase in mean arterial pressure, and often, a *slowing* of heart rate. Bradycardia, the reflex compensatory response elicited by this agent, is the *exact opposite* of the drug's direct action; yet it is completely predictable if the integration of cardiovascular function by the ANS is understood.

#### B. Presynaptic Regulation

The principle of negative feedback control is also found at the presynaptic level of autonomic function. Important presynaptic feedback inhibitory control mechanisms have been shown to exist at most nerve endings. A well-documented mechanism involves the  $\alpha_2$  receptor located on noradrenergic nerve terminals. This receptor is activated by norepinephrine and similar molecules; activation diminishes further release of norepinephrine from these nerve endings (Table 6-4). The mechanism of this G protein-mediated effect involves inhibition of the inward calcium current that causes vesicular fusion and transmitter release. Conversely, a presynaptic  $\beta$  receptor appears to facilitate the release of norepinephrine from some adrenergic neurons. Presynaptic receptors that respond to the primary transmitter substance released by the nerve ending are called **autoreceptors**. Autoreceptors are usually inhibitory, but in addition to the excitatory  $\beta$  receptors on noradrenergic fibers, many cholinergic fibers, especially somatic motor fibers, have excitatory nicotinic autoreceptors.

Control of transmitter release is not limited to modulation by the transmitter itself. Nerve terminals also carry regulatory receptors that respond to many other substances. Such **heteroreceptors** may be activated by substances released from other nerve terminals that synapse with the nerve ending. For example, some vagal fibers in the myocardium synapse on sympathetic noradrenergic nerve terminals and inhibit norepinephrine release. Alternatively, the ligands for these receptors may diffuse to the receptors from the blood or from nearby tissues. Some of the transmitters and receptors identified to date are listed in Table 6-4. Presynaptic

TABLE 6-2 Direct effects of autonomic drugs on various organs	
Organ	Direct effects
Ileum	Smooth muscle
Intestine	Smooth muscle
Ciliary muscle	Contraction
Heart	Stimulation
Sinus node	Excitation
Esophagus	Smooth muscle
Contractility	Smooth muscle
Blood vessels	Smooth muscle
Skin	Smooth muscle
Skeletal muscle vessels	Smooth muscle
Endothelium of vessels in heart, brain, lungs	Smooth muscle
Bronchial smooth muscle	Smooth muscle
Gastrointestinal tract	Smooth muscle
Smooth muscle	Smooth muscle
Walls	Smooth muscle
Sphincters	Smooth muscle
Secretion	Smooth muscle
Gastrointestinal smooth muscle	Smooth muscle
Bladder wall	Smooth muscle
Sphincter	Smooth muscle
Uterus, pregnant	Smooth muscle
Pelvic splanchnic nerves	Smooth muscle
Skin	Smooth muscle
Hypothalamic smooth muscle	Smooth muscle
Sweat glands	Smooth muscle
Emotion	Smooth muscle
Adaptive processes	Smooth muscle
Metabolic functions	Smooth muscle
Liver	Smooth muscle
Arterioles	Smooth muscle
Basal metabolism	Smooth muscle

Note: Direct effects are drawn in bold. If a drug acts on a tissue, but does not have a direct effect, it is not included in this table. The inhibition of most drugs results in inhibition of the performance of the organ. Conversely, the stimulation of most drugs results in stimulation of the performance of the organ. The inhibition of smooth muscle results in relaxation of the smooth muscle. Conversely, the stimulation of smooth muscle results in contraction of the smooth muscle.

Figure 6: Raw Photo of Page 100 of Book 1

#### 14. THE SONG OF THE SEA WIND

How it sings, sings, sings,  
Blowing sharply from the sea-line,  
With an edge of salt that stings;  
How it laughs aloud, and passes,  
As it cuts the close cliff-grasses:  
How it sings again, and whistles  
As it shakes the stout sea-thistles—  
How it sings!

How it shrieks, shrieks, shrieks,  
In the crannies of the headland,  
In the gashes of the creeks;  
How it shrieks once more, and catches  
Up the yellow foam in patches;  
How it hurls it out and over  
To the corn-field and the clover—  
How it shrieks!

How it roars, roars, roars,  
In the iron under-caverns,  
In the hollows of the shores;  
How it roars anew, and thunders,  
As the strong hull splits and sunders:  
And the spent ship, tempest-driven,  
On the reef lies rent and riven—  
How it roars!

How it wails, wails, wails,  
In the tangle of the wreckage,  
In the flapping of the sails;  
How it sobs away, subsiding,  
Like a tired child after chiding;  
And across the ground-swell rolling  
You can hear the bell-buoy tolling—  
How it wails!

—AUSTIN DOBSON.

## References

2022. *Using the Tesseract OCR Engine in R*. <https://cran.r-project.org/web/packages/tesseract/vignettes/intro.html>.
- Bache, Stefan Milton, and Hadley Wickham. 2020. *Magrittr: A Forward-Pipe Operator for r*. <https://CRAN.R-project.org/package=magrittr>.
- Haahr, Mads. 2019. “RANDOM.ORG - True Random Number Service.” *Random.org*. <https://www.random.org/>.
- Inc, SeatGeek. 2014. *fuzzywuzzy: Fuzzy String Matching in Python*. <https://github.com/seatgeek/fuzzywuzzy>.
- Katzung, Betram G., and Bertram G. Katzung. 2018. *Basic & Clinical Pharmacology / by Betram g. Katzung*. McGraw-Hill.
- Mouselimis, Lampros. 2021. *fuzzywuzzyR: Fuzzy String Matching*. <https://CRAN.R-project.org/package=fuzzywuzzyR>.
- Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. <https://CRAN.R-project.org/package=here>.
- Muschelli, John. 2017. *Diffr: Display Differences Between Two Files Using Codediff Library*. <https://CRAN.R-project.org/package=diffr>.
- Ooms, Jeroen. 2017. “Tesseract and Magick: High Quality OCR in r: R-Bloggers.” *R*. <https://www.r-bloggers.com/2017/08/tesseract-and-magick-high-quality-ocr-in-r/>.
- . 2021. *Magick: Advanced Graphics and Image-Processing in r*. <https://CRAN.R-project.org/package=magick>.
- . 2022. *Tesseract: Open Source OCR Engine*. <https://CRAN.R-project.org/package=tesseract>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. <http://www.rstudio.com/>.
- Sariyar, Murat, and Andreas Borg. 2022. *RecordLinkage: Record Linkage Functions for Linking and Duplicating Data Sets*. <https://CRAN.R-project.org/package=RecordLinkage>.
- Tanselle, G. Thomas, and Martin C. Battestin. 1993. *The Life and Work of Fredson Bowers*. Bibliographical Society of the University of Virginia.
- “The History of OCR, Optical Character Recognition : Schantz, Herbert f : Free Download, Borrow, and Streaming.” 1982. *Internet Archive*. [Manchester Center, Vt.] : Recognition Technologies Users Association. <https://archive.org/details/historyofocropti0000scha/page/n5/mode/2up>.
- Ushey, Kevin, JJ Allaire, and Yuan Tang. 2022. *Reticulate: Interface to ‘Python’*. <https://CRAN.R-project.org/package=reticulate>.
- Wickham, Hadley. 2007. “Reshaping Data with the reshape Package.” *Journal of Statistical Software* 21 (12): 1–20. <http://www.jstatsoft.org/v21/i12/>.
- . 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Xie, Yihui. 2014. “Knitr: A Comprehensive Tool for Reproducible Research in R.” In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. <http://www.crcpress.com/product/isbn/9781466561595>.
- . 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. <https://yihui.org/knitr/>.
- . 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://yihui.org/knitr/>.
- Zhu, Hao. 2021. *kableExtra: Construct Complex Table with ‘Kable’ and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.