

# The Brain Tumor Segmentation - Metastases (BraTS-METS) Challenge 2023: Brain Metastasis Segmentation on Pre-treatment MRI

Ahmed W. Moawad <sup>1,\*</sup>,<sup>α,β</sup> Anastasia Janas <sup>2,\*</sup>,<sup>α,β,δ</sup> Ujjwal Baid <sup>3,\*</sup>,<sup>α,β</sup> Divya Ramakrishnan <sup>2,\*</sup>,<sup>α,β</sup> Rachit Saluja <sup>4,5,\*</sup>,<sup>α,β,γ</sup> Nader Ashraf <sup>6,7,\*</sup>,<sup>α,β,γ</sup> Nazanin Maleki <sup>2,6,\*</sup>,<sup>α,β,δ</sup> Leon Jekel <sup>8,\*</sup>,<sup>α,β</sup> Nikolay Yordanov <sup>9,δ,κ</sup> Pascal Fehringer <sup>10,δ,κ</sup> Athanasios Gkampenesis <sup>11,δ,κ</sup> Raisa Amiruddin <sup>6,α,δ</sup> Amirreza Manteghinejad <sup>6,α</sup> Maruf Adewole <sup>12,α</sup> Jake Albrecht <sup>13,α</sup> Udunna Anazodo <sup>12,14,α</sup> Sanjay Aneja <sup>15,α</sup> Syed Muhammad Anwar <sup>16,α</sup> Timothy Bergquist <sup>17,α</sup> Veronica Chiang <sup>18,α</sup> Verena Chung <sup>13,α</sup> Gian Marco Conte <sup>17,α</sup> Farouk Dako <sup>19,α</sup> James Eddy <sup>13,α</sup> Ivan Ezhov <sup>20,α</sup> Nastaran Khalili <sup>21,α</sup> Keyvan Farahani <sup>22,α</sup> Juan Eugenio Iglesias <sup>23,α</sup> Zhifan Jiang <sup>24,α</sup> Elaine Johanson <sup>25,α</sup> Anahita Fathi Kazerooni <sup>21,26,27,α</sup> Florian Kofler <sup>28,α</sup> Kiril Krantchev <sup>2,α,β,ε,δ</sup> Dominic LaBella <sup>29,α</sup> Koen Van Leemput <sup>30,α</sup> Hongwei Bran Li <sup>23,α</sup> Marius George Linguraru <sup>16,31,α</sup> Xinyang Liu <sup>24,α</sup> Zeke Meier <sup>32,α</sup> Bjoern H Menze <sup>33,α</sup> Harrison Moy <sup>2,α,β,ε</sup> Klara Osenberg <sup>2,α,β</sup> Marie Piraud <sup>34,α</sup> Zachary Reitman <sup>29,α</sup> Russell Takeshi Shinozawa <sup>35,α</sup> Chunhao Wang <sup>29,α</sup> Benedikt Wiestler <sup>28,α</sup> Walter Wiggins <sup>36,α</sup> Umber Shafique <sup>37,α,η</sup> Klara Willms <sup>2,β</sup> Arman Avesta <sup>2,38,β</sup> Khaled Bousabarah <sup>39,β,ε</sup> Satrajit Chakrabarty <sup>40,41,β</sup> Nicolo Gennaro <sup>42,β</sup> Wolfgang Holler <sup>39,β,ε</sup> Manpreet Kaur <sup>43,β,ε</sup> Pamela LaMontagne <sup>44,β</sup> MingDe Lin <sup>45,β,ε</sup> Jan Lost <sup>46,β,ε</sup> Daniel S. Marcus <sup>44,β</sup> Ryan Maresca <sup>15,β,ε</sup> Sarah Merkaj <sup>47,β,ε</sup> Gabriel Cassinelli Pedersen <sup>48,β,ε</sup> Marc von Reppert <sup>49,β,ε</sup> Aristeidis Sotiras <sup>44,50,β</sup> Oleg Teytelboym <sup>1,β</sup> Niklas Tillmans <sup>51,β,ε</sup> Malte Westerhoff <sup>39,β,ε</sup> Ayda Youssef <sup>52,β</sup> Devon Godfrey <sup>29,β</sup> Scott Floyd <sup>29,β</sup> Andreas Rauschecker <sup>53,β</sup> Javier Villanueva-Meyer <sup>53,β</sup> Irada Pflüger <sup>54,β</sup> Jaeyoung Cho <sup>54,β</sup> Martin Bendeszus <sup>54,β</sup> Gianluca Brugnara <sup>54,β</sup> Justin Cramer <sup>55,η</sup> Gloria J. Guzman Perez-Carillo <sup>56,η</sup> Derek R. Johnson <sup>17,η</sup> Anthony Kam <sup>57,η</sup> Benjamin Yin Ming Kwan <sup>58,η</sup> Lillian Lai <sup>59,η</sup> Neil U. Lall <sup>60,η</sup> Fatima Memon <sup>61,62,63,η</sup> Mark Krycia <sup>61,η</sup> Satya Narayana Patro <sup>64,η</sup> Bojan Petrovic <sup>65,η</sup> Tiffany Y. So <sup>66,η</sup> Gerard Thompson <sup>67,68,η</sup> Lei Wu <sup>69,η</sup> E. Brooke Schrickel <sup>70,η</sup> Anu Bansal <sup>71,θ</sup> Frederik Barkhof <sup>72,73,θ</sup> Cristina Besada <sup>74,θ</sup> Sammy Chu <sup>69,θ</sup> Jason Druzgal <sup>75,θ</sup> Alexandru Dusoi <sup>76,θ</sup> Luciano Farage <sup>77,θ</sup> Fabricio Feltrin <sup>78,θ</sup> Amy Fong <sup>79,θ</sup> Steve H. Fung <sup>80,θ</sup> R. Ian Gray <sup>81,θ</sup> Ichiro Ikuta <sup>55,θ</sup> Michael Iv <sup>82,θ</sup> Alida A. Postma <sup>83,84,θ</sup> Amit Mahajan <sup>2,θ</sup> David Joyner <sup>75,θ</sup> Chase Krumpelman <sup>42,θ</sup> Laurent Letourneau-Guillon <sup>85,θ</sup> Christie M. Lincoln <sup>86,θ</sup> Mate E. Maros <sup>87,θ</sup> Elka Miller <sup>88,θ</sup> Fanny Morón <sup>89,θ</sup> Esther A. Nimchinsky <sup>90,θ</sup> Ozkan Ozsarlar <sup>91,θ</sup> Uresh Patel <sup>92,θ</sup> Saurabh Rohatgi <sup>38,θ</sup> Atin Saha <sup>93,94,θ</sup> Anousheh Sayah <sup>95,θ</sup> Eric D. Schwartz <sup>96,97,θ</sup> Robert Shih <sup>98,θ</sup> Mark S. Shiroishi <sup>99,θ</sup> Juan E. Small <sup>100,θ</sup> Manoj Tanwar <sup>101,θ</sup> Jewels Valerie <sup>102,θ</sup> Brent D. Weinberg <sup>103,θ</sup> Matthew L. White <sup>104,θ</sup> Robert Young <sup>93,θ</sup> Vahe M. Zohrabian <sup>105,θ</sup> Aynur Azizova <sup>106,θ</sup> Melanie Maria Theresa Brüßeler <sup>43,κ</sup> Mohanad Ghonim <sup>107,κ</sup> Mohamed Ghonim <sup>107,κ</sup> Abdullah Okar <sup>108,κ</sup> Luca Pasquini <sup>93,κ</sup> Yasaman Sharifi <sup>109,κ</sup> Gagandeep Singh <sup>110,κ</sup> Nico Sollmann <sup>111,112,113,κ</sup> Theodora Soumala <sup>11,κ</sup> Mahsa Taherzadeh <sup>114,κ</sup> Philipp Vollmuth <sup>54,115,β,γ</sup> Martha Folty-Dumitru <sup>54,β,γ</sup> Ajay Malhotra <sup>2,β,γ</sup> Aly H. Abayazeed <sup>82,γ</sup> Francesco Dellepiane <sup>116,γ</sup> Philipp Lohmann <sup>117,118,γ</sup> Víctor M. Pérez-García <sup>119,γ</sup> Hesham Elhalawani <sup>120,γ</sup> Maria Correia de Verdier <sup>121,122,γ</sup> Sanaria Al-Rubaiey <sup>123,λ</sup> Rui Duarte Armindo <sup>124,λ</sup> Kholod Ashraf <sup>52,λ</sup> Moamen M. Asla <sup>125,λ</sup> Mohamed Badawy <sup>126,λ</sup> Jeroen Bisschop <sup>127,λ</sup> Nima Broomand Lomer <sup>128,λ</sup> Jan Bukatz <sup>123,λ</sup> Jim Chen <sup>129,λ</sup> Petra Cimflova <sup>130,λ</sup> Felix Corr <sup>131,λ</sup> Alexis Crawley <sup>132,λ</sup> Lisa Deptula <sup>133,λ</sup> Tasneem Elakhdar <sup>52,λ</sup> Islam H. Shawali <sup>52,λ</sup> Shahriar Faghani <sup>17,λ</sup> Alexandra Frick <sup>134,λ</sup> Vaibhav Gulati <sup>135,λ</sup> Muhammad Ammar Haider <sup>136,λ</sup> Fátima Hierro <sup>137,λ</sup> Rasmus Holmboe Dahl <sup>138,λ</sup> Sarah Maria Jacobs <sup>139,λ</sup> Kuang-chun Jim Hsieh <sup>89,λ</sup> Sedat G. Kandemirli <sup>59,λ</sup> Katharina Kersting <sup>123,λ</sup> Laura Kida <sup>123,λ</sup> Sofia Kollia <sup>140,λ</sup> Ioannis Koukoulithras <sup>141,λ</sup> Xiao Li <sup>103,λ</sup> Ahmed Abouelatta <sup>52,λ</sup> Aya Mansour <sup>52,λ</sup> Ruxandra-Catrinel Maria-Zamfirescu <sup>123,λ</sup> Marcela Marsiglia <sup>142,λ</sup> Yohana Sarahi Mateo-Camacho <sup>143,λ</sup> Mark McArthur <sup>144,λ</sup> Olivia McDonnell <sup>145,λ</sup> Maire McHugh <sup>146,λ</sup> Mana Moassefi <sup>147,λ</sup> Samah Mostafa Morsi <sup>86,λ</sup> Alexander Munteanu <sup>148,λ</sup> Khanak K. Nandolia <sup>149,λ</sup> Syed Raza Naqvi <sup>150,λ</sup> Yalda Nikanpour <sup>151,λ</sup> Mostafa Alnoury <sup>152,λ</sup> Abdullah Mohamed Aly Nouh <sup>153,λ</sup> Francesca Pappafava <sup>154,λ</sup> Markand D. Patel <sup>155,λ</sup> Samantha Petrucci <sup>53,λ</sup> Eric Rawie <sup>156,λ</sup> Scott Raymond <sup>157,λ</sup> Borna Roohani <sup>108,λ</sup> Sadeq Sabouhi <sup>158,λ</sup> Laura M. Sanchez-Garcia <sup>159,λ</sup> Zoe Shaked <sup>123,λ</sup> Pokhraj P. Suthar <sup>160,λ</sup> Talissa Altes <sup>161,λ</sup> Edvin Isufi <sup>161,λ</sup> Yaseen Dhemesh <sup>162,λ</sup> Jaime Gass <sup>161,λ</sup> Jonathan Thacker <sup>161,λ</sup> Abdul Rahman Tarabishy <sup>163,λ</sup> Benjamin Turner <sup>164,λ</sup> Sebastiano Vacca <sup>165,λ</sup> George K. Vilanilam <sup>164,λ</sup> Daniel Warren <sup>162,λ</sup> David Weiss <sup>166,λ</sup> Fikadu Worede <sup>6,λ</sup> Sara Yousry <sup>52,λ</sup> Wondwossen Lerebo <sup>6,μ</sup> Alejandro Aristizabal <sup>167,168,π</sup> Alexandros Karargyris <sup>167,π</sup> Hasan Kassem <sup>167,π</sup> Sarthak Pati <sup>3,167,169,π</sup> Micah Sheller <sup>167,170,π</sup> Katherine E. Link <sup>171,α,β</sup> Evan Calabrese <sup>172,α,β</sup> Nourel hoda Tahon <sup>161,α,β</sup> Ayman Nada <sup>161,α,β</sup> Yuri S. Velichko <sup>42,α,β</sup> Spyridon Bakas <sup>3,37,173,α,β,φ</sup> Jeffrey D. Rudie <sup>122,174,α,β,η,φ</sup> Mariam Aboian <sup>6,α,β,η,φ,†</sup>

<sup>1</sup> Trinity health Mid Atlantic Hospitals, Darby, PA, USA

<sup>2</sup> Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA

<sup>3</sup> Division of Computational Pathology, Department of Pathology and Laboratory Medicine, School of Medicine, Indiana University, Indianapolis, IN, USA

- 4** Department of Electrical and Computer Engineering, Cornell University and Cornell Tech, New York, NY, USA  
**5** Department of Radiology, Weill Cornell Medicine, New York, NY, USA  
**6** Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, PA, USA  
**7** College of Medicine, Alfaisal University, Riyadh, Saudi Arabia  
**8** DKFZ Division of Translational Neurooncology at the WTZ, German Cancer Consortium, DKTG Partner Site, University Hospital Essen, Essen, Germany  
**9** Faculty of Medicine, Medical University - Sofia, Sofia, Bulgaria  
**10** Faculty of Medicine, Jena University Hospital, Friedrich Schiller University Jena, Jena, Germany  
**11** University of Ioannina School of Medicine, Ioannina, Greece  
**12** Medical Artificial Intelligence Lab, Crestview Radiology, Lagos, Nigeria  
**13** Sage Bionetworks, Seattle, WA, USA  
**14** Montreal Neurological Institute, McGill University, Montreal, Canada  
**15** Department of Therapeutic Radiology, Yale School of Medicine, New Haven, CT, USA  
**16** Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, D.C., USA  
**17** Department of Radiology, Mayo Clinic, Rochester, MN, USA  
**18** Department of Neurosurgery, Yale School of Medicine, New Haven, CT, USA  
**19** Center for Global Health, Perelman School of Medicine, University of Pennsylvania, PA, USA  
**20** Department of Informatics, Technical University Munich, Germany  
**21** Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA  
**22** Cancer Imaging Program, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA  
**23** Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA  
**24** Children's National Hospital, Washington, D.C., USA  
**25** PrecisionFDA, U.S. Food and Drug Administration, Silver Spring, MD, USA  
**26** Department of Neurosurgery, University of Pennsylvania, Philadelphia, PA, USA  
**27** Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, USA  
**28** Department of Neuroradiology, Technical University of Munich, Munich, Germany  
**29** Department of Radiation Oncology, Duke University Medical Center, Durham, NC, USA  
**30** Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark  
**31** Departments of Radiology and Pediatrics, George Washington University School of Medicine and Health Sciences, Washington, D.C., USA  
**32** Booz Allen Hamilton, McLean, VA, USA  
**33** Biomedical Image Analysis & Machine Learning, Department of Quantitative Biomedicine, University of Zurich, Switzerland  
**34** Helmholtz AI, Helmholtz Munich, Germany  
**35** Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA, USA  
**36** Duke University School of Medicine, Durham, NC, USA  
**37** Department of Radiology and Imaging Sciences, Indiana University, Indianapolis, IN, USA  
**38** Department of Radiology, Neuroradiology, Massachusetts General Hospital, Boston, MA, USA  
**39** Visage Imaging, GmbH, Berlin, Germany  
**40** Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO, USA  
**41** GE HealthCare, San Ramon, CA, USA  
**42** Department of Radiology, Northwestern University, Feinberg School of Medicine, Chicago, IL, USA  
**43** Ludwig Maximilian University, Munich, Germany  
**44** Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO, USA  
**45** Visage Imaging, Inc, San Diego, CA, USA  
**46** Department of Neurosurgery, Heinrich-Heine University, Moorenstrasse 5, Dusseldorf, Germany  
**47** University of Ulm, Ulm, Germany **48** University of Göttingen, Göttingen, Germany **49** University of Leipzig, Leipzig, Germany **50** Institute for Informatics, Data Science & Biostatistics, Washington University School of Medicine, St. Louis, MO, USA  
**51** Department of Diagnostic and Interventional Radiology, Medical Faculty, University Dusseldorf, Dusseldorf, Germany  
**52** Cairo University, Cairo, Egypt  
**53** Department of Radiology and Biomedical Imaging, University of California San Francisco, CA, USA  
**54** Department of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany

- 55** Department of Radiology, Mayo Clinic, Phoenix, AZ, USA  
**56** Neuroradiology Section, Mallinckrodt Institute of Radiology, Washington University in St. Louis, St. Louis, MO, USA  
**57** Loyola University Medical Center, Hines, IL, USA  
**58** Department of Radiology, Queen's University, Kingston, ON, Canada  
**59** Department of Radiology, University of Iowa Hospitals and Clinics, Iowa City, IA, USA  
**60** Children's Healthcare of Atlanta, GA, USA  
**61** Carolina Radiology Associates, Myrtle Beach, SC, USA  
**62** McLeod Regional Medical Center, Florence, SC, USA  
**63** Medical University of South Carolina, Charleston, SC, USA  
**64** University of Arkansas Medical Center, Little Rock, AR, USA  
**65** NorthShore Endeavor Health, Evanston, IL, USA  
**66** Department of Imaging and Interventional Radiology, The Chinese University of Hong Kong, Hong Kong SAR  
**67** Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom  
**68** Department of Clinical Neurosciences, NHS Lothian, Edinburgh, United Kingdom  
**69** Department of Radiology, University of Washington, Seattle, WA, USA  
**70** Department of Radiology, Ohio State University College of Medicine, Columbus, OH, USA  
**71** Albert Einstein Medical Center, Hartford, CT, USA  
**72** Amsterdam UMC, location Vrije Universiteit, Netherlands  
**73** University College London, United Kingdom  
**74** Hospital Italiano de Buenos Aires, Buenos Aires, Argentina  
**75** Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, Virginia, USA  
**76** Klinikum Hochrhein, Waldshut-Tiengen, Germany  
**77** Centro Universitario Euro-American (UNIEURO), Brasília, DF, Brazil  
**78** Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX, USA  
**79** Southern District Health Board, Dunedin, New Zealand  
**80** Department of Radiology, Houston Methodist, Houston, TX, USA  
**81** University of Tennessee Medical Center, Knoxville, TN, USA  
**82** Department of Radiology, Stanford University, Stanford, CA, USA  
**83** Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, Maastricht, the Netherlands  
**84** Mental Health and Neuroscience Research Institute, Maastricht University, Maastricht, the Netherlands  
**85** Centre Hospitalier de l'Université de Montréal and Centre de Recherche du CHUM Montréal, Canada  
**86** Department of Neuroradiology, MD Anderson Cancer Center, Houston, TX, USA  
**87** Departments of Neuroradiology & Biomedical Informatics, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany  
**88** Department of Diagnostic and Interventional Radiology, SickKids Hospital, University of Toronto, Canada  
**89** Department of Radiology, Baylor College of Medicine, Houston, TX, USA  
**90** Department of Radiology, New Jersey Medical School, Newark, NJ, USA  
**91** Department of Radiology, AZ Monica, Antwerp Area, Belgium  
**92** Medicolegal Imaging Experts LLC, Mercer Island, WA, USA  
**93** Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA  
**94** Weill Cornell Medical College, New York, NY, USA  
**95** MedStar Georgetown University Hospital, Washington, D.C., USA  
**96** Department of Radiology, St. Elizabeth's Medical Center, Boston, MA, USA  
**97** Department of Radiology, Tufts University School of Medicine, Boston, MA, USA  
**98** Walter Reed National Military Medical Center, Bethesda, MD, USA  
**99** Keck School of Medicine, Los Angeles, CA, USA  
**100** Lahey Hospital and Medical Center, Burlington, MA, USA  
**101** Department of Radiology, University of Alabama, Birmingham, AL, USA  
**102** Department of Radiology, University of North Carolina School of Medicine, Chapel Hill, NC, USA  
**103** Department of Radiology and Imaging Sciences, Emory University, Atlanta, GA, USA  
**104** University of Nebraska Medical Center, Omaha, NE, USA  
**105** Northwell Health, Zucker Hofstra School of Medicine at Northwell, North Shore University Hospital, Hempstead, New York, NY, USA

- 106** Cancer Center Amsterdam, Imaging and Biomarkers, Amsterdam, The Netherlands  
**107** Department of Radiology, Ain Shams University, Cairo, Egypt  
**108** University of Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany  
**109** Department of Radiology, Iran University of Medical Sciences, Tehran, Iran  
**110** Columbia University Irving Medical Center, New York, NY, USA  
**111** Department of Diagnostic and Interventional Radiology, University Hospital Ulm, Ulm, Germany  
**112** Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany  
**113** TUM-Neuroimaging Center, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany  
**114** Department of Radiology, Arad Hospital, Tehran, Iran  
**115** Department of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany  
**116** Functional and Interventional Neuroradiology Unit, Bambino Gesù Children's Hospital, Rome, Italy  
**117** Institute of Neuroscience and Medicine (INM-4), Research Center Juelich, Juelich, Germany  
**118** Department of Nuclear Medicine, University Hospital RWTH Aachen, Aachen, Germany  
**119** Mathematical Oncology Laboratory & Department of Mathematics, University of Castilla-La Mancha, Spain  
**120** Department of Radiation Oncology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA  
**121** Department of Surgical Sciences, Section of Neuroradiology, Uppsala University, Sweden  
**122** Department of Radiology, University of California San Diego, CA, USA  
**123** Charité-Universitätsmedizin Berlin (Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health), Berlin, Germany  
**124** Department of Neuroradiology, Western Lisbon Hospital Centre (CHLO), Portugal  
**125** Zagazig University, Zagazig, Egypt  
**126** Diagnostic Radiology Department, Wayne State University, Detroit, MI  
**127** Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland.  
**128** Faculty of Medicine, Guilan University of Medical Sciences, Rasht, Iran  
**129** Department of Radiology/Division of Neuroradiology, San Diego Veterans Administration Medical Center/UC San Diego Health System, San Diego, CA, USA  
**130** Department of Radiology, University of Calgary, Calgary, Canada  
**131** EDU Institute of Higher Education, Villa Bighi, Chaplain's House, Kalkara, Malta  
**132** Bay Imaging Consultants, Walnut Creek, CA, USA  
**133** Ross University School of Medicine, Bridgetown, Barbados  
**134** Department of Neurosurgery, Vivantes Klinikum Neukölln, Berlin, Germany  
**135** Mercy Catholic Medical Center, Darby, PA, USA  
**136** C.M.H. Lahore Medical College, Lahore, Pakistan  
**137** Neuroradiology Department, Pedro Hispano Hospital, Matosinhos, Portugal  
**138** Department of Radiology, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark  
**139** Rijnstate Hospital, Arnhem, Netherlands **140** National and Kapodistrian University of Athens, School of Medicine, Athens, Greece  
**141** Department of Neurosurgery, University Hospital of Ioannina, Ioannina, Greece  
**142** Department of Radiology, Brigham and Women's Hospital, Massachusetts General Hospital, Boston, MA, USA  
**143** Department of Neuroradiology, Universidad Autónoma de Nuevo León, México  
**144** Department of Radiological Sciences, University of California Los Angeles, Los Angeles, CA, USA  
**145** Gold Coast University Hospital, Queensland Health, Australia  
**146** Department of Radiology Manchester NHS Foundation Trust, North West School of Radiology, Manchester, United Kingdom  
**147** Artificial Intelligence Lab, Department of Radiology, Mayo Clinic, Rochester, MN, USA  
**148** Corewell Health West, MI, USA  
**149** Department of Radiodiagnosis, All India Institute of Medical Sciences Rishikesh, India  
**150** Windsor Regional Hospital, Western University, Ontario, Canada  
**151** Artificial Intelligence & Informatics, Mayo Clinic, Rochester, MN, USA  
**152** Department of Radiology, University of Pennsylvania, PA, USA  
**153** Department of Radiology, Life Care Hospital, Freetown, Sierra Leone

- 154** Department of Medicine and Surgery, Università degli Studi di Perugia, Italy  
**155** Department of Neuroradiology, Imperial College Healthcare NHS Trust, London, United Kingdom  
**156** Department of Radiology, Michigan Medicine, Ann Arbor, MI, USA  
**157** Department of Radiology, University of Vermont Medical Center, Burlington, VT, USA  
**158** Isfahan University of Medical Sciences, Isfahan, Iran  
**159** Department of Radiology, The American British Cowdray Medical Center, Mexico City, Mexico  
**160** Rush University Medical Center, Chicago, IL, USA  
**161** Radiology Department, University of Missouri, Columbia, MO, USA  
**162** Washington University School of Medicine in St. Louis, St. Louis, MO, USA  
**163** Department of NeuroRadiology, Rockefeller Neuroscience Institute, West Virginia University. Morgantown, WV, USA  
**164** Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom  
**165** University of Cagliari, School of Medicine and Surgery, Cagliari, Italy  
**166** Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany  
**167** MLCommons, San Francisco, CA, USA  
**168** Factored, Palo Alto, CA, USA  
**169** Center For Federated Learning in Medicine, Indiana University, Indianapolis, IN, USA  
**170** Intel Corporation, Hillsboro, OR, USA  
**171** New York University School of Medicine, New York, NY, USA  
**172** Department of Radiology, Duke University Medical Center, Durham, NC, USA  
**173** Department of Neurological Surgery, School of Medicine, Indiana University, Indianapolis, IN, USA  
**174** Department of Radiology, Scripps Clinic Medical Group, CA, USA
- \* Equal First Authors  
α Organizer  
β Data contributors  
γ BraTS2024 Organizer  
δ International lead  
ε Annotator  
η Super Approver  
θ Rest of Approvers  
κ Super Annotator  
λ Rest of Annotators  
μ Stats  
π MLCommons  
ϕ Equal Senior Authors  
† Corresponding Author — aboianm@chop.edu

## Abstract

The translation of AI-generated brain metastases (BM) segmentation into clinical practice relies heavily on diverse, high-quality annotated medical imaging datasets. The BraTS-METS 2023 challenge has gained momentum for testing and benchmarking algorithms using rigorously annotated internationally compiled real-world datasets. This study presents the results of the segmentation challenge and characterizes the challenging cases that impacted the performance of the winning algorithms. Untreated brain metastases on standard anatomic MRI sequences (T1, T2, FLAIR, T1PG) from eight contributed international datasets were annotated in stepwise method: published UNET algorithms, student, neuroradiologist, final approver neuroradiologist. Segmentations were ranked based on lesion-wise Dice and Hausdorff distance (HD95) scores. False positives (FP) and false negatives (FN) were rigorously penalized, receiving a score of 0 for Dice and a fixed penalty of 374 for HD95. The mean scores for the teams were calculated. Eight datasets comprising 1303 studies were annotated, with 402 studies (3076 lesions) released on Synapse as publicly available datasets to challenge competitors. Additionally, 31 studies (139 lesions) were held out for validation, and 59 studies (218 lesions) were used for testing. Segmentation accuracy was measured as rank across subjects, with the winning team achieving a LesionWise mean score of 7.9. The Dice score for the winning team was  $0.65 \pm 0.25$ . Common errors among the leading teams included false negatives for small lesions and misregistration of masks in space. The Dice scores and

lesion detection rates of all algorithms diminished with decreasing tumor size, particularly for tumors smaller than 100 mm<sup>3</sup>. In conclusion, algorithms for BM segmentation require further refinement to balance high sensitivity in lesion detection with the minimization of false positives and negatives. The BraTS-METS 2023 challenge successfully curated well-annotated, diverse datasets and identified common errors, facilitating the translation of BM segmentation across varied clinical environments and providing personalized volumetric reports to patients undergoing BM treatment.

**Keywords**

BraTS, BraTS-METS, Medical image analysis challenge, Brain metastasis, Brain tumor segmentation, Machine learning, Artificial Intelligence

**Article informations**

©2024 BraTS-METS Team. License: CC-BY 4.0

## 1. Introduction

**B**rain metastases represent the most common malignancy affecting the adult central nervous system (Le Rhun et al., 2021), affecting an estimated 20–40% of patients with systemic cancer (Percy et al., 1972; Tabouret et al., 2012; Posner, 1978; Nayak et al., 2012). Patients commonly have multiple lesions at different stages of treatment, therefore radiologic evaluation often extends beyond a mere comparison with the most recent scan. In clinical practice, a comprehensive assessment frequently involves reviewing several previous scans to monitor the progression or changes in the metastases over time which can be laborious and time-consuming (Jekel et al., 2022b; Kaur et al., 2023; Cassinelli Petersen et al., 2022).

The shift toward automated volumetric analysis and lesion organization in evaluating BMs is a transformative (Kaur et al., 2023; Ocaña-Tienda et al., 2023), transcending the conventional qualitative assessment methods to a personalized and time-efficient approach. Artificial intelligence (AI) based volumetric BMs assessments will not only improve the precision of measurements but also provide high-quality personalized reports of individual treatment response of brain metastases and thus influence patient outcomes; it's about democratizing access to high-quality care Pinto-Coelho, 2023; Najjar, 2023; Tang, 2019. By integrating automated volumetric analysis into clinical practice, we can ensure more reliable and consistent measurements, extending these advanced diagnostic capabilities beyond specialized centers to a broader range of healthcare settings. Improved accessibility of personalized reporting is crucial, particularly for patients in regions where such specialized services were previously unavailable, thus broadening the scope of quality care to include more comprehensive and timely monitoring of disease progression and response to treatment.

The intricate task of accurately detecting, segmenting, and assessing BMs is pivotal for devising effective therapeutic strategies and prognostication. However, the efficacy of machine learning algorithms in this realm is inherently tied to the availability and quality of annotated medical imaging datasets (Zhou et al., 2020; Zhang et al., 2020; Xue et al., 2020; Jeong et al., 2024; Grøvik et al., 2020; Dikici et al., 2020, 2022; Charron et al., 2018; Bousabarah et al., 2020). Historically, the scarcity of large-scale, annotated datasets in the medical imaging field has limited the potential of machine learning algorithms. Many researchers find themselves constrained to smaller, local institutional datasets, which limits algorithm generalizability across different institutions (Greenspan et al., 2016). In this context, medical image analysis challenges—competitions to establish accurate segmentation algorithms—have emerged as crucial platforms, facilitating the development, testing, and bench-

marking of machine learning algorithms by providing access to extensive, meticulously labeled, multi-center, real-world datasets.

Specifically, the domain of BMs analysis stands to benefit immensely from such collaborative initiatives. The complexities associated with BMs, such as the variability in size, shape, and location of lesions, necessitate sophisticated machine learning approaches that can adapt to the diverse characteristics of these metastatic manifestations (Cho et al., 2021). Moreover, the dynamic nature of BMs, with changes occurring over time and in response to treatment, underscores the need for algorithms capable of longitudinal assessment and multi-lesion segmentation.

The 2023 Brain Tumor Segmentation - Metastases (BraTS-METS) challenge marked a significant shift from previous BraTS challenges, which centered on adult brain diffuse astrocytoma (Zhang et al., 2020; Xue et al., 2020; Jeong et al., 2024). The scope was broadened to encompass a variety of brain tumor entities, thereby addressing the issue of data scarcity and methodological complexities inherent in earlier challenges. This challenge prioritized the segmentation of BMs on pre-treatment MR imaging. The goal of BraTS-METS 2023 was to establish a robust, accurate algorithm for segmenting metastatic lesions of virtually any size on diagnostic magnetic resonance imaging (MRI) using T1-weighted (T1) pre-contrast, T1 post-contrast, T2-weighted (T2), and fluid attenuated inversion recovery (FLAIR) sequences. The resulting standardized auto-segmentation algorithm was made openly accessible, thus facilitating its integration into clinical and research protocols across institutions.

Initially, the intention was to develop an algorithm dedicated to segmenting pre-treatment BMs (Figure 1, Step 1). This algorithm was fine-tuned to delineate the enhancing tumor, peritumoral edema, and necrotic portions of the metastases (Figure 1, Step 2). The ultimate aim was to establish a BMs consortium for future collaborative research (Figure 1, Step 3). This consortium is designed to foster a collaborative research environment, not only for the development of BM imaging algorithms but also for their clinical translation and community education efforts.

## 2. Background

Standard-of-care for evaluation of BMs includes qualitative assessment of changes in lesion size and number and two dimensional measurements performed by radiologists manually on PACS workstation. In clinical trials, the Response Assessment in Neuro-Oncology Brain Metastases (RANO-BM) guidelines predominantly rely on measuring the unidimensional longest diameter of lesions (Lin et al., 2015). However, these traditional criteria may not fully capture the complex dynamics and morphological changes

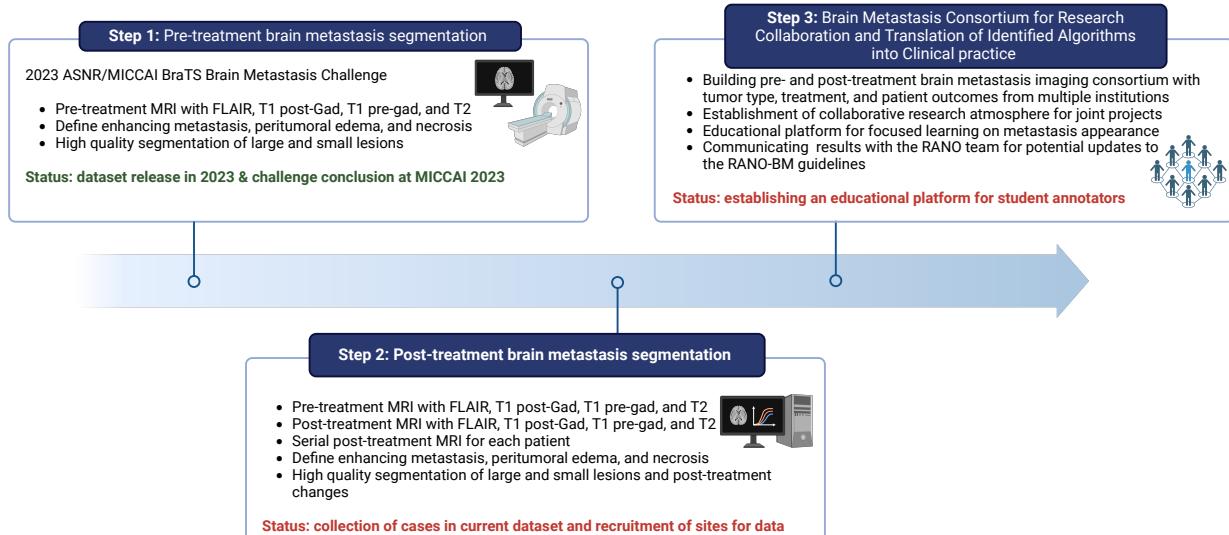


Figure 1: Flow chart outlining the BraTS-METS 2023 vision, beginning with the pre-treatment BMs segmentation during the 2023 ASNR/MICCAI BraTS challenge. In this phase, segmentations were conducted on a select dataset subset to refine the dataset for algorithm development by participants. The dataset is set to expand in subsequent challenges through ongoing annotation of contributed brain MRIs. Future challenges will incorporate datasets with annotated post-treatment BMs, segmentations including the hemorrhagic component of tumors, and non-skull-stripped images to enhance the evaluation of dural-based and osseous metastases. These datasets, coupled with clinical data and patient demographics, will contribute to an inter-institutional BMs consortium, fostering collaborative research and the clinical application of algorithms through partnerships between academia and industry.

of BMs over time, particularly given the heterogeneity and irregular growth patterns often associated with these lesions.

Recent advances in MRI technology, particularly the adoption of high-resolution 3D sequences such as T1 magnetization prepared rapid acquisition gradient-echo, T1 fast spoiled gradient-echo, and T1 three-dimension high-resolution inversion recovery-prepared fast spoiled gradient-recalled, have significantly enhanced our ability to detect and monitor smaller BMs. The traditional threshold for target lesions, as outlined in the RANO-BM criteria proposed by Lin et al., set the minimum size at 10 mm in longest diameter, visible on two or more axial slices with a 5 mm or less interval (Lin et al., 2015). However, with the advancements in imaging, lesions as small as 1-2 mm can now be reliably detected, but because of significant inter-rater variability in measurement of lesions smaller than 5 mm, the consensus criteria still requires a lesion of at least 10 mm to be considered as measurable disease. Introduction of improved reproducibility and low variability between algorithm based measurements provides a potential for future re-evaluation of standardized assessment criteria to include smaller lesions. Indeed, recent practices have seen a shift towards a 5 mm minimum size threshold, aligning with the capabilities of current MRI technology, as highlighted by Qian et al. (2017).

Integration of automated techniques, such as deep learn-

ing algorithms for segmentation and assessment, offers a promising avenue approach to enhance the precision and efficiency of volumetric evaluations, aligning with the requirements of the RANO-BM guidelines (Kanakarajan et al., 2023; Wang et al., 2023a; Yoo et al., 2022). The importance of multi-lesional segmentation and continuous assessment across serial imaging cannot be overstated. Such a comprehensive approach can benefit from the integration of automatic algorithms that are capable of efficiently detecting and segmenting metastases across multiple imaging time points, including pre- and post-treatment scans. The enhanced precision and efficiency of clinical assessments can complement the expertise of radiologists and other clinicians, which would aid not only in tracking disease progression and response to treatment but also in identifying new lesions at the earliest possible stage.

Despite the potential benefits, the routine implementation of such automated techniques in clinical settings faces significant hurdles, given the extensive time required and the variability inherent in imaging techniques across different temporal scans. This variability often arises from disparate imaging equipment and the fact that different radiologists may interpret sequential scans for a single patient differently, introducing acquisition heterogeneity and inter-reader variability (Buchner et al., 2023; Mi et al., 2020).

Addressing the detection and segmentation challenges

associated with smaller BMs is therefore of paramount importance. The successful development of targeted algorithms will expedite their translation to and adoption in clinical practice, providing a vital resource in the management of BMs. By successfully overcoming those challenges, we can provide algorithms that can be readily translated and implemented in clinical settings.

### 3. Related Works

While challenges remain in the field of automated BMs segmentation, recent studies are indicative of a promising trajectory toward achieving high levels of automation, consistency, and adaptability in clinical practice (Jekel et al., 2022b; Kanakarajan et al., 2023; Dang et al., 2022; Jekel et al., 2022a; Chen et al., 2023b). Kanakarajan et al. (2023) demonstrated a significant advancement with their development of a fully automated segmentation method for BMs using T1 contrast-enhanced MR images, which could significantly aid in evaluating treatment effects post-stereotactic radiosurgery. Similarly, Buchner et al. (2023) have identified core MRI sequences that are essential for reliable automatic BMs segmentation, providing a foundation for standardized imaging protocols and enhancing algorithmic consistency across various clinical settings.

The integration of multi-phase delayed enhanced MR images has been explored by Chen et al. (2023b), who reported improvements in the accuracy of both segmentation and classification of BMs. This approach addressed the critical need for refined diagnostic tools that can adapt to the complex nature of BMs. Furthermore, Ottesen et al. (2023) have extended the capabilities of deep learning algorithms by implementing 2.5D and 3D segmentation techniques on multinational MRI data, enhancing the robustness and adaptability of these systems for diverse clinical environments.

The ongoing development and refinement of these automated segmentation tools are set to revolutionize the way BMs are assessed, bringing about a significant enhancement in the consistency and quality of patient care (Jekel et al., 2022b; Jalalifar et al., 2023). Yoo et al. (2023) underscored the importance of the data domain in self-supervised learning for accurate BMs detection and segmentation. This development points toward the creation of more adaptable and robust systems capable of functioning effectively across a variety of clinical scenarios. Moreover, advancements in the reduction of false positives within automated BMs segmentation underscore the growing feasibility and effectiveness of these technologies, even in diverse clinical environments, cementing their role as invaluable assets in medical imaging (Ghesu et al., 2022; Liew et al., 2023; Ziyaei et al., 2023).

Detecting smaller metastatic lesions, typically ranging

from 1 to 2 mm, is pivotal in patient prognosis and treatment planning. Given the increased reliance on SRS (Vogelbaum et al., 2022), accurately identifying the exact number and localization of these small metastases becomes even more critical to ensure effective treatment and minimize the risk of missed targets, which could necessitate additional interventions, cause treatment delays, and increase healthcare costs (Minniti et al., 2011; Schnurman et al., 2022; Chen et al., 2023c). The gross total volume (GTV) of BMs is potentially a critical prognostic indicator, yet its clinical utility remains largely untapped due to the absence of validated volumetric segmentation tools. The considerable effort required to detect and volumetrically segment all lesions, irrespective of size, poses a significant challenge. While existing glioma-focused segmentation algorithms, such as those developed by Applied Computer Vision Lab & Division of Medical Image Computing, Germany, have shown promising accuracy for larger metastases as measured by Dice scores, their efficacy diminishes with smaller lesions.

Efforts to release publicly available BM datasets have varied significantly in their criteria and quality, contributing to inconsistencies in algorithm training and validation. Table 1 provides a summary of previously publicly available datasets.

The development of a universally accepted, metastasis-specific AI tool represents a considerable gap in the current landscape, posing a barrier to the standard clinical use of GTV assessment for prognostication in patients with BMs. This challenge is compounded by the lack of a comprehensive public dataset, which would facilitate a fair comparison of existing BMs segmentation models. The availability of such a dataset could significantly accelerate progress by enabling researchers to benchmark and refine their models against a standardized dataset, thereby enhancing the reliability and accuracy of AI-powered segmentation tools. Bridging these gaps is essential for advancing the integration of AI in the prognostic evaluation of BMs, ultimately improving patient management and treatment outcomes.

## 4. Materials & Methods

### 4.1 Data

The BraTS-METS dataset included retrospectively collected multiparametric MRI (mpMRI) scans from diverse institutions, representing the variability in imaging protocols and equipment reflective of global clinical practices. Inclusion criteria encompassed MRI scans with the presence of untreated BMs with T1 pre-contrast, T1 post-contrast, T2, and FLAIR sequences. Participating institutions had obtained Institutional Review Board and Data Transfer Agreement approvals before contributing data, ensuring compliance with regulatory standards. These scans were

Table 1: Overview of publicly available datasets for BMs.

Public Dataset	Data Publisher	Number of case	Difference from BraTS datasets
NYUMets (Oermann et al., 2023)	New York University	1,429 patients	<ul style="list-style-type: none"> <li>Contains post therapy cases</li> <li>Not all patients have images</li> <li>Most cases without segmented BM</li> </ul>
BrainMetShare (Grøvik et al., 2020)	Stanford University	156 patients	<ul style="list-style-type: none"> <li>Does not contain T2 sequence</li> <li>Contains post therapy cases</li> <li>Only contains TC subregion</li> <li>Available in JPEG format</li> </ul>
UCSF-BMSR (Rudie et al., 2024)	University of California San Francisco	412 patients	<ul style="list-style-type: none"> <li>Contains synthetic T2 images</li> <li>Contains post therapy cases</li> </ul>
Brain-TR-GammaKnife (Wang et al., 2023b)	University of Mississippi	47 patients	<ul style="list-style-type: none"> <li>Does not contain T2 images</li> <li>Contains post therapy cases</li> </ul>
MOLAB (Ocaña-Tienda et al., 2023)	University of Castilla-La Mancha	75 patients	<ul style="list-style-type: none"> <li>Contains post therapy cases</li> <li>Recently published</li> <li>Not all BMs are segmented</li> </ul>

then centralized and curated for consistency.

Exclusion criteria included the presence of prior treatment changes, lack of one of the required MRI sequences, or imaging not technically acceptable due to motion or other significant imaging artifacts. The cases where post-treatment changes were noted were reserved for BraTS-METS 2024.

The dataset allocation for the BraTS-METS 2023 challenge adhered to the standard machine learning protocol, with 70% designated for training, 10% for validation, and 20% for testing. Ground truth (GT) labels were provided exclusively for the training set, while the validation set remained unlabeled to ensure integrity in algorithmic evaluation. The testing set was kept hidden from the participants. The use of additional data, whether public or private, was restricted to prevent bias in the algorithmic ranking process. Participants were allowed to reference external datasets only for publication purposes and were required to disclose

such usage transparently in their manuscripts, along with results derived from the BraTS-METS 2023 dataset.

#### 4.2 Imaging Data Description

The mpMRI scans included four sequences: non-enhanced T1, post-gadolinium-contrast T1 (T1Gd), T2, and non-enhanced T2-FLAIR, procured from various scanners and protocols. Standardized pre-processing was applied to all the BraTS-METS mpMRI scans. Specifically, the applied pre-processing routines included conversion of the DICOM files to the NIfTI file format, co-registration to the same anatomical template (SRI24)(Rohlfing et al., 2010), resampling to a uniform isotropic resolution ( $1\text{mm}^3$ ), and, finally, skull stripping (Isensee et al., 2019). The pre-processing pipeline was made publicly available through the Cancer Imaging Phenomics Toolkit (CaPTk) (Pati et al., 2020; Rathore et al., 2018) and the Federated Tumor Segmenta-

tion (FeTS) tool (Pati et al., 2022). Conversion to Neuroimaging Informatics Technology Initiative (NIftI) stripped the accompanying metadata from the Digital Imaging and Communications in Medicine (DICOM) images and removed all protected health information from the DICOM headers. Furthermore, skull stripping mitigated potential facial reconstruction/recognition of the patient (Greenspan et al., 2016; Cho et al., 2021). The specific approach used for skull stripping was based on a novel deep learning approach that accounts for the brain shape prior and was agnostic to the MRI sequence input (Juluru et al., 2020; Schwarz et al., 2019).

#### 4.3 Tumor Labels

The annotation of tumor sub-regions aligned with Visually AcceSable Rembrandt Images (VASARI) feature visibility and encompassed three labels: Gd-enhancing tumor (ET - label 3), surrounding non-enhancing FLAIR hyperintensity (SNFH - label 2), and the non-enhancing tumor core (NETC – label 1). ET is described as the enhancing portion of the tumor, characterized by areas of hyperintensity in T1Gd that are brighter than T1. NETC is identified as the presumed necrotic core of the tumor, which is evident as a non-enhancing focus surrounded by enhancing tumor. SNFH is defined as the peritumoral edema and tumor infiltrated tissue, indicated by the abnormal hyperintense signal on the T2-FLAIR images, which includes the infiltrative non-enhancing tumor, as well as vasogenic edema in the peritumoral region. In previous BraTS challenges, ET was segmented as label 4. However, starting from BraTS 2023, ET has been segmented as label 3 for consistency. The sub-regions are shown in Figure 2.

#### 4.4 Tumor Annotation Protocol

The BraTS initiative, in consultation with domain experts, defined various tumor sub-regions to provide a standardized approach for their assessment and evaluation. However, alternative criteria for delineation could be established, resulting in slightly different tumor sub-regions. To ensure consistency in the GT delineations across various annotators, the following tumor annotation protocol was designed. Structural mpMRI volumes were considered (T1, T1Gd, T2, T2-FLAIR).

The BraTS-METS 2023 challenge focuses on three regions of interest:

1. Whole Tumor (WT) = Label 1 + Label 2 + Label 3
2. Tumor Core (TC) = Label 1 + Label 3
3. Enhancing Tumor (ET) = Label 3

WT describes the complete extent of the disease, encompassing TC and the peritumoral edematous/invasive

tissue, typically depicted by the abnormal hyper-intense signal in the T2-FLAIR volume. While the radiologic definition of tumor boundaries, especially in infiltrative tumors such as gliomas, presents a well-known challenge, this is less problematic in BMs, which typically have well-defined borders of the contrast-enhancing portion. In most cases, the boundaries of the contrast-enhancing region of the BM and the surrounding FLAIR hyperintense edema are well defined. One of the major challenges in segmenting BMs lies in the overlap of edema between multiple lesions, which is why the segmentation of ET is separated from WT and treated as distinct entities.

#### 4.5 Annotation Pipeline

To ensure uniformity in data imaging and tumor labeling, we established a comprehensive annotation pipeline (Figure 3). This pipeline facilitates the development of accurate GT labels and is divided into five key stages: pre-segmentation, annotation refinement, technical quality control (QC), initial approval, and final approval.

#### 4.6 Pre-segmentation

The initial phase involved pre-segmenting imaging volumes using three distinct approaches:

1. nnU-Net trained on the University of California, San Francisco BMs Stereotactic Radiosurgery (UCSF-BMSR) MRI Dataset (Rudie et al., 2024), which creates the ET label and was fused with predictions of NETC and SNFH from an nnU-Net trained on the pre-treatment BraTS 2021 glioma dataset.
2. nnU-Net trained on AURORA multicenter study (Kaur et al., 2023), which creates SNFH and tumor core (ET + NETC) labels.
3. nnU-Net trained on Heidelberg University Hospital dataset (Pflüger et al., 2022), which creates SNFH and tumor core labels.

The label fusion process varied for each label. SNFH (label – 2) was fused using the STAPLE fusion algorithm to aggregate the segmentations from each automated segmentation algorithm, accounting for systematic errors (Warfield et al., 2004). ET (label – 3) was fused using the minority voting algorithm to aggregate all enhancing tumor voxels identified by the automated segmentation algorithms, due to varying accuracies in detecting small metastases. NETC (label – 1) is only produced by the nnU-Net trained on UCSF-BMSR. Algorithms trained on AURORA and Heidelberg datasets only segment TC and SNFH. Therefore, NETC overlays both ET and SNFH labels.

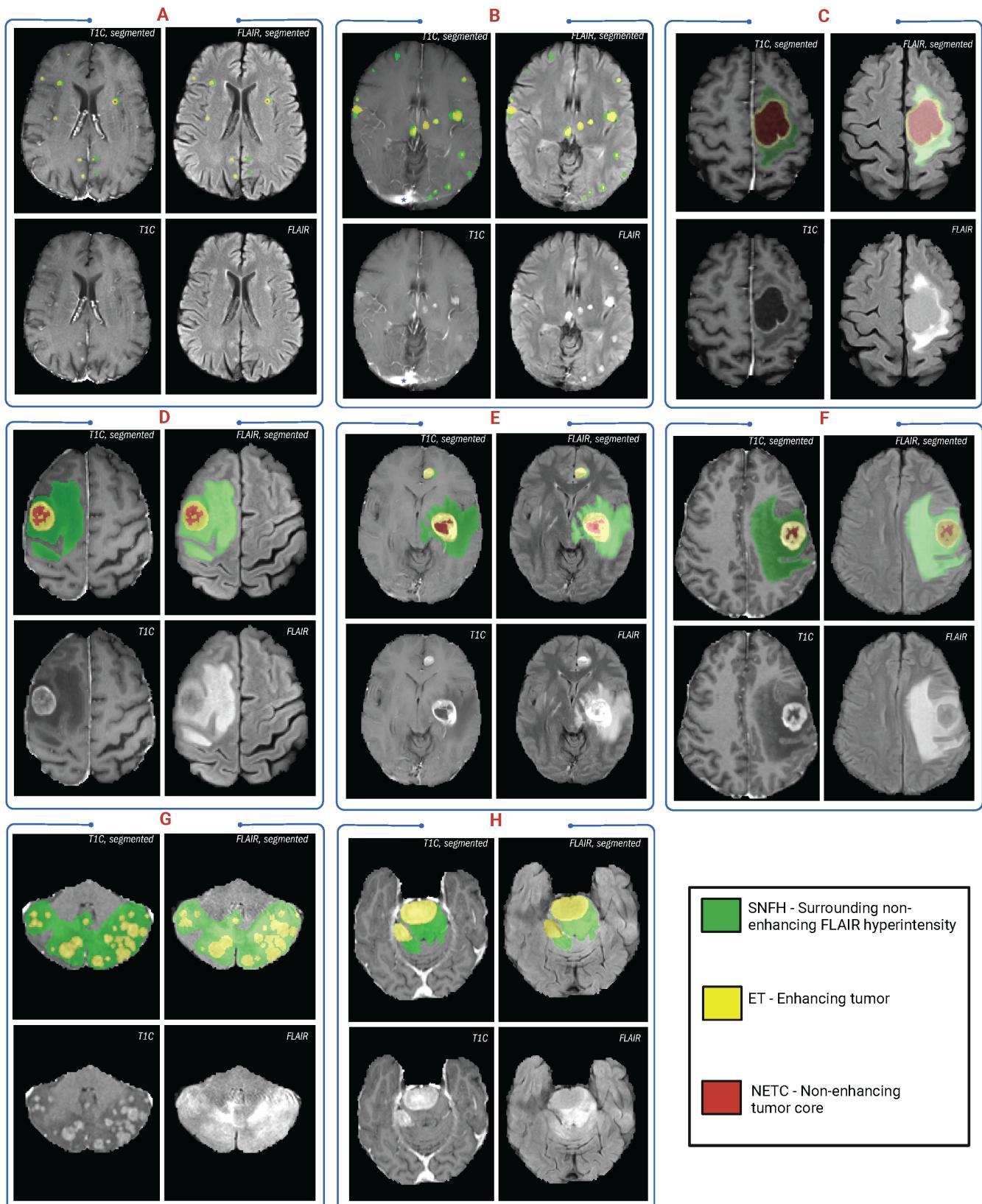


Figure 2: Image panels illustrating the annotated tumor sub-regions across various mpMRI scans with segmentations of ET (yellow), SNFH (green), and NETC (red) done on ITK-SNAP.

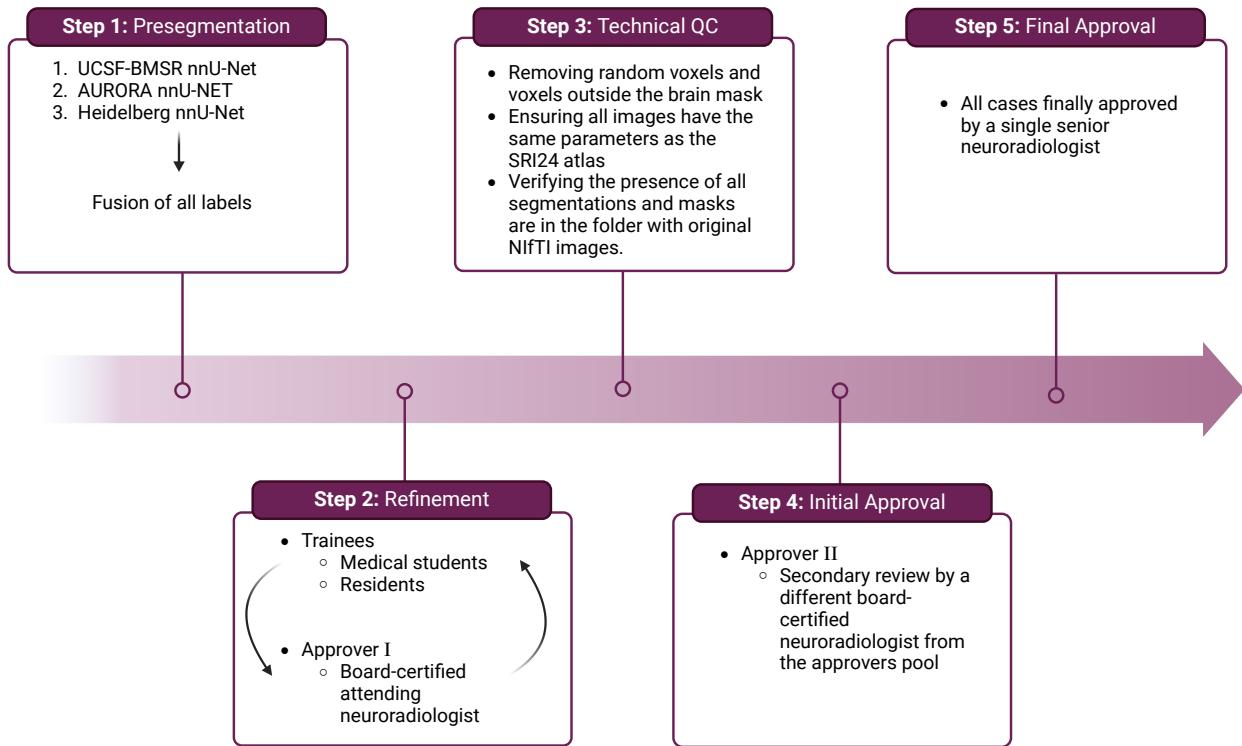


Figure 3: BraTS-METS 2023 annotation pipeline.

#### 4.7 Annotation Refinement and Initial Approval

All pre-segmentations from the three models, along with fused segmentations, were provided to the annotators. Subtraction images, in which the non-contrast T1 sequence is digitally subtracted from the post-contrast T1 sequence, were also provided to aid in the annotation refinement process. Annotations were performed by a diverse group of more than 150 student annotators and volunteer neuroradiology experts, under the supervision of annotator coordinators (A.J. and K.K.). Cases requiring re-annotation due to incompleteness were identified and returned for correction. During the process of annotation, the trainees participated in group reviews of cases, asked questions, and attended lectures by expert imagers. Completed student annotations were then reviewed by a pool of 52 experienced board-certified attending neuroradiologists (approvers) recruited by the American Society of Neuroradiology, ensuring quality control and uniformity with the SRI24 atlas standards.

Approvers reviewed the volunteer annotations and either approved the case or returned it to students for re-annotation. Additionally, a QC process was implemented, which included removing all random voxels and any voxels outside the brain mask, ensuring all images had the same parameters (space, orientation, and origin) as the SRI24 atlas, and verifying the presence of all segmentations and segmentation masks are in the folder with original NIfTI images.

#### 4.8 Annotation Final Approval

Following refinement, each case underwent a secondary review by a different board-certified neuroradiologist from the approver pool, ensuring accurate metastasis segmentation and adherence to inclusion criteria. In cases of discrepancy, the second approvers made the necessary changes themselves without reverting to the trainees. Finally, a neuroradiologist (M.A.) with over 6 years of brain tumor expertise conducted a final dataset review, guaranteeing consistency across all annotations.

#### 4.9 Common Errors of Automated Segmentations

Based on observations from previous BraTS challenges, common errors in automated segmentations were identified. The most typical errors in the current challenge included:

1. Automated algorithms missing small metastases. Enhancing metastasis was fused using the minority voting algorithm to aggregate all enhancing tumor voxels identified by the three algorithms. However, many small metastases were missed and were manually segmented by neuroradiology attendings.
2. Segmentation of white matter changes from microvascular disease. Peritumoral edema segmentations were checked by neuroradiology attendings and modified.
3. The segmentation of non-enhancing lesions that have intrinsic T1 hyperintensity. Voxels with intrinsic T1

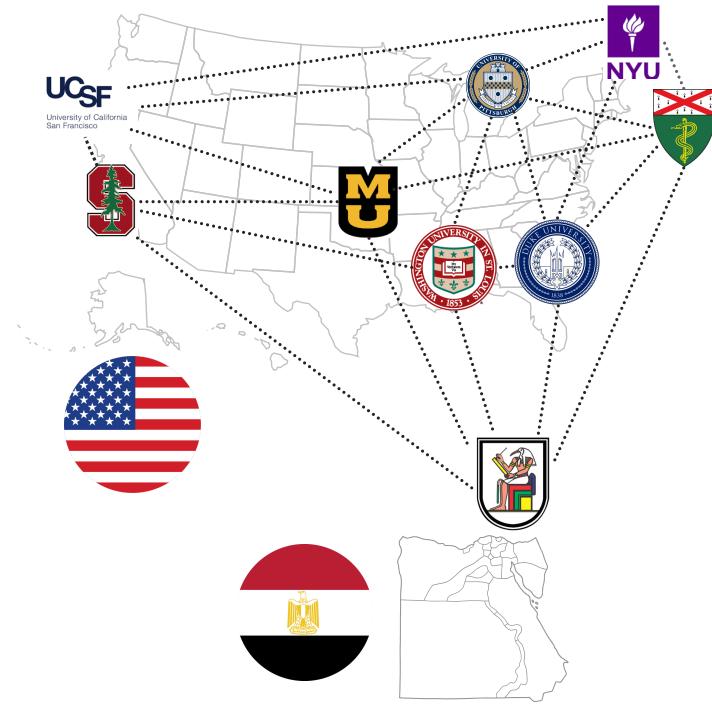


Figure 4: Map of institutions that expressed interest in contributing data to the BraTS-METS challenge.

hyperintensity were manually removed from ET segmentations.

These insights led to specific adjustments in the annotation process to enhance accuracy.

#### 4.10 Performance Evaluation Framework

Participants were offered a baseline approach implemented in the Generally Nuanced Deep Learning Framework (GaNDLF), a modular open-source framework maintained by the MLCommons organization. GaNDLF provides popular network architectures, but also allows users to leverage the functionality of other libraries, such as PILLOW and MONAI. Submissions were packaged in MLCube containers as described in the instructions provided in the Synapse platform. These submissions were registered to MLCommons' MedPerf, an open federated AI/ML evaluation platform. MedPerf automated the pipeline of running the participants' models on the evaluation datasets of each contributing site's data and calculating evaluation metrics on the resulting predictions. Finally, the Synapse platform retrieved the metrics results from the MedPerf server and ranked them to determine the winner.

Performance evaluation was based on Dice scores and 95% Hausdorff distance (HD95) for individual segmented lesions as defined by the three regions of interest: ET, TC and WT. Given that BMs are often small, sometimes

comprising only a few voxels, it was clinically significant to assess segmentation algorithms based on their capacity to accurately detect and delineate both small and large lesions. Teams were ranked based on a combination of lesionwise Dice and Hausdorff distance scores across all evaluated test cases. False positives and false negatives were rigorously penalized, receiving a score of 0 for Dice and a fixed penalty of 374 for HD95. This methodical approach was uniformly applied across the three designated tissue classes, with subsequent aggregation of results by taking the mean score for each CaselD within each tissue category.

$$\text{Lesion-wise Dice Score} = \frac{\sum_i^L \text{Dice}(l_i)}{TP + FN + FP} \quad (1)$$

$$\text{Lesion-wise HD95} = \frac{\sum_i^L \text{HD95}(l_i)}{TP + FN + FP} \quad (2)$$

where  $L$  is the total number of GT lesions and  $TP$ ,  $FP$ ,  $FN$  are the number of true positive, false positive and false negative lesions respectively.

All participants were evaluated and ranked using the same unseen testing data, which was not accessible to them. They were required to upload their containerized method to the evaluation platforms. The final top-ranked teams were announced at the 2023 Medical Image Computing and Computer Assisted Intervention Society (MICCAI) annual

meeting, with monetary prizes awarded to the top-ranked teams in both tasks of the challenge.

For this challenge, each team was ranked relative to its competitors for each of the testing subjects, for each evaluated region (i.e., ET, TC, WT), and for each measure (i.e., Dice and Hausdorff). For example, each team was ranked for 59 subjects, for 3 regions, and for 2 metrics, which resulted in  $59 \times 3 \times 2 = 354$  individual rankings. The final ranking score (FRS) for each team was then calculated by first averaging across all these individual rankings for each patient (i.e., cumulative rank), and then averaging these cumulative ranks across all patients for each participating team. This ranking scheme has also been adopted in other challenges with satisfactory results, such as the Ischemic Stroke Lesion Segmentation challenge (Maier et al., 2017).

We then conducted further permutation testing to determine statistical significance of the relative rankings between each pair of teams. This permutation testing reflected differences in performance that exceeded those that might be expected by chance. Specifically, for each team, we started with a list of observed subject-level cumulative ranks, i.e., the actual ranking described above. For each pair of teams, we repeatedly randomly permuted (i.e., for 100,000 times) the cumulative ranks for each subject. For each permutation, we calculated the difference in the FRS between this pair of teams. The proportion of times the difference in FRS calculated using randomly permuted data exceeded the observed difference in FRS (i.e., using the actual data) indicated the statistical significance of their relative rankings as a p-value. These values were reported in an upper triangular matrix, providing insights of statistically significant differences across each pair of participating teams.

#### 4.11 Analysis

The competition framework encompassed evaluations across three key regions: ET, TC, and WT, utilizing two primary metrics: lesion-wise Dice and lesion-wise HD95. These metrics have been developed primarily to evaluate the performance of models at the level of individual lesions, rather than on a whole-image basis. This approach ensured that our evaluation did not favor models that only captured large lesions, a limitation commonly observed with standard Dice scores. By assessing models on a lesion-by-lesion basis, we gained insights into their ability to segment all sizes of BMs accurately.

To implement this evaluation framework, we first isolated the lesion tissues (i.e., ET, TC, WT). We applied dilation to the GT labels for WT, TC, and ET to gauge the lesion's extent. This technique ensured that during connected component analysis, small lesions adjacent to a primary lesion were not misclassified as separate entities. It is crucial to note that the GT labels remained unchanged

throughout this process. We conducted a 26-connectivity connected component analysis on the predicted labels and compared each component to the corresponding GT label on a component-by-component basis. We calculated the Dice scores and HD95 scores individually for each lesion (or component), assigning the aforementioned penalty, to all false positives and negatives. Subsequently, we computed the mean score for each specific case.

Acknowledging the variability in lesion significance arising due to human error, a volumetric threshold of 2 voxels ( $2 \text{ mm}^3$ ) was established by an expert panel of clinical radiologists, below which the models' performance on deemed "small/false" lesions is not considered in the evaluation. This approach was primarily adopted to ensure that participants were not unfairly penalized for stray voxels in the GT labels, which may result from human error, or for small lesions unrelated to the pathology central to the challenge. The expert panel of clinical radiologists also determined the dilation factor, which was uniformly applied for combining lesions in the GT masks. A dilation factor of 1 voxel in 3D space was chosen because BMs can be small, and it is important to avoid combining these small BMs.

The code and detailed information on the lesion-wise evaluation metrics can be found here<sup>1</sup>.

#### 4.12 Dataset

Multiple datasets were contributed by individual institutions and were in various stages of annotation and approval (Figure 4).

### 5. Results

#### 5.1 Dataset Sources

Our annotation and approval pipeline, as previously described, was applied to datasets from a variety of institutions, including New York University (NYU), Yale University, Washington University, Cairo University (CairoU), Duke University, and the University of Missouri. The annotated NYU dataset is uniquely hosted on the NYU website (access to the data can be requested by filling the form)<sup>2</sup>, separate from the public BraTS repository. As for the UCSF dataset, synthetic T2 images were generated and shared on the UCSF website<sup>3</sup>. The Stanford University dataset, despite being publicly available, was not incorporated into our primary dataset due to the lack of T2 image sequences. These datasets were available and optional for additional training. For logistical reasons, the UCSF, Stanford, and NYU datasets were excluded from the validation and test phases of our project.

1. <https://github.com/rachitsaluja/BraTS-2023-Metrics>

2. <https://nyumets.org/>; <https://forms.gle/UqE6VMgCtpT21rmu7>

3. <https://imagingdatasets.ucsf.edu/dataset/1>

Table 2: Dataset sources in the BraTS-METS 2023 challenge. In the training dataset, 474 cases from UCSF and Stanford were included as optional because they did not have original T2 weighted images.

Dataset Source	Total cases reviewed	Excluded	Training	Validation	Test
Duke	37	0	26	4	7
CairoU	45	10	32	1	2
Missouri	25	3	16	2	4
WashU	40	1	27	4	8
Yale	225	30	137	20	38
NYU*	221	57	164	0	0
UCSF <sup>^</sup>	560	236	324	0	0
Stanford <sup>^</sup>	150	0	150	0	0
Total	1,303	337	402 (474 optional)	31	59

\* The NYU dataset is part of the official challenge. Because it is hosted on a separate website, it is not included in the validation or test set.

<sup>^</sup> UCSF and Stanford datasets are not part of the official challenge. Both datasets are provided as optional training sets.

Table 3: Lesion count and sizes for each dataset group.

Dataset Group	ET lesion-count (total)	ET lesion-count median (IQR)	ET lesion-size median (IQR)	WT lesion-count (total)	WT lesion-count median (IQR)	WT lesion-size median (IQR)
Training* (n = 402)	3076	3 (7)	65 (287)	2618	3 (5)	121 (804)
Validation (n = 31)	139	3 (4)	141 (664)	119	3(3)	591 (3318)
Testing (n = 59)	218	2 (3)	132 (613)	193	2 (3)	322 (8624)

\* The training group does not include the optional UCSF and Stanford datasets.

In all, 2712 cases were received from various institutes of which 1303 cases were reviewed from eight institutions. After 337 cases were excluded, 876 cases were allocated into the training (n = 402; UCSF and Stanford datasets cases that were optional, n = 474), validation (n = 31), and testing (n = 59) groups (Table 2). All the source institutions were located in the United States, except for one in Egypt.

## 5.2 Lesion Characteristics

Table 3 provides a detailed overview of lesion count and sizes across the different dataset groups used in the BraTS-METS 2023 challenge. These data demonstrate the variation in lesion count and size across the dataset groups.

## 5.3 Performance Analysis

Table 4 provides the relative ranking for each team. Team NVAUTO ranked first in the challenge, with an average rank across subjects of 7.9 and a PatientWise mean of 0.38. Team SY placed second with a PatientWise mean of 0.41 across all patients. The supplementary material depicts the pitfall cases with figures illustrating the false positives or missed lesions.

Figure 5 provides a patient-wise comparison of segmentation accuracy across the different participating teams. The boxplots reflect the distribution of each team's accuracy per patient case per lesion—across all cases within the test dataset, with lower value signifying better performance. The teams NVAUTO, SY, and blackbean showed a notably higher median accuracy, alongside a relatively narrow in-

Table 4: Top-performing teams ranking with cumulative ranks across subjects. Lower scores indicate better performance.

Team Name	Cumulative ranks across subjects	Lesion-wise mean	Rank
NVAUTO	466	7.9	1
SY	503	8.5	2
blackbean	571.5	9.7	3
CNMCPMI2023	689	11.7	4
isahajmistry	817	13.8	5
DeepRadOnc	907.5	15.4	6
MIASINTEF	1002	17	7

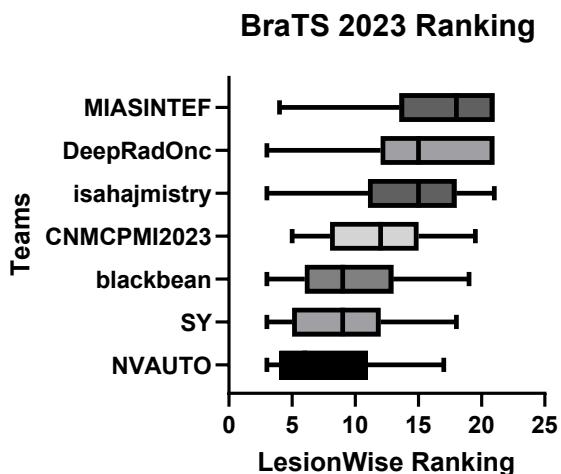


Figure 5: BraTS-METS 2023 boxplots of LesionWise ranking across patients for all participating teams on the BraTS 2023 test set (lower is better).

terquartile range (IQR). Conversely, DeepRadOnc displayed a wider IQR.

A description of the algorithms used by the top four winning teams are shown in Table 5.

#### 5.4 Detailed Performance by Tumor Entities

Table 6 delineates the comparative performance of each participating team's Dice scores for each tumor entity (i.e., ET, TC, and WT). The team NVAUTO secured the top rank across all categories, exhibiting a mean Dice score of 0.60 for ET, 0.65 for TC, and 0.62 for WT. Notably, SY and blackbean shared the second rank in the ET segmentation, with a mean of 0.57. Figures 6, 7, and 8 further highlight the lesion-wise Dice scores (shown as panels A) and HD95 (shown as panels B) for each participating team for each tumor entity.

Figure 9 illustrates a comparative evaluation across the three tumor regions of interest where performance of the segmentation models is quantified using three metrics: lesion detection rate, sensitivity, and positive predictive value

(PPV). The lesion detection rate was led by NVAUTO with rates of 76% for ET, 78% for TC, and 80% for WT. Closely following were blackbean and SY, with both achieving a 75% detection rate for ET and TC, and 76% and 72% for WT, respectively. In terms of sensitivity, NVAUTO again showed superior performance, with 90% for ET, 91% for TC, and 90% for WT, reflecting a high true positive rate. blackbean and SY exhibited comparably high sensitivity, around 89–90% across tumor entities. PPV results depicted NVAUTO at the forefront with 82% for ET, 84% for TC, and 84% for WT. Following suit, blackbean maintained a PPV of 79% across all tumor entities, and SY showcased a slightly lower yet robust PPV performance with 76%.

#### 5.5 Algorithm Sensitivity to Lesion Size

Figure 10 provides insight into the models' performance in segmenting lesions of different sizes. This was analyzed by calculating a running average within an expanding window of tumor volume, starting with only the smallest tumors and progressively including larger lesions (Kelahan et al., 2022).

The graphs collectively indicate that segmentation algorithm performance diminishes as tumor size decreases, with all teams facing challenges in maintaining high Dice scores and lesion detection rates for smaller tumors. The HD95 data suggest that algorithms struggled with precision in delineating the contours of smaller lesions, reflected in greater distances from the ground truth, a trend particularly noticeable for tumors less than 100 mm<sup>3</sup> in volume. Despite these challenges, NVAUTO consistently outperformed its counterparts.

## 6. Discussion

The use of machine learning in medical imaging has brought notable improvements in detecting and segmenting BMs. Clinical evaluation of BMs has unique complexity because it requires volumetric measurements and organization of lesions to provide granular details on individual lesion treat-

Table 5: Description of algorithms used by the top 4 winning teams.

Team Name & DL algorithm	Description
NVAUTO (SegResNet from MONAI Auto3DSeg)	<ul style="list-style-type: none"> <li>▪ MONAI native (uses transforms, loaders, losses, networks components of MONAI)</li> <li>▪ 4-channel input, which is a concatenation of four different MRI scans</li> <li>▪ Input data is normalized to have zero mean and unit standard deviation for each channel.</li> <li>▪ Employs random cropping to a fixed size of 224x224x144 pixels</li> <li>▪ AdamW optimizer with a learning rate of 2e-4 is used in combination with a cosine annealing scheduler</li> <li>▪ Model is trained for a range of 300 to 1000 epochs, using 5-fold cross-validation</li> <li>▪ A combined Dice-Focal loss function is utilized for training</li> <li>▪ Data augmentation techniques include spatial transformations (random rotations, scaling, flips) and intensity modifications (random adjustments to intensity/contrast, addition of noise, and blur)</li> <li>▪ Code reference: GitHub - MONAI and SegResNetDS</li> </ul>
SY (3D TransUNet Model (Chen et al., 2023a))	<ul style="list-style-type: none"> <li>▪ 3D nnUNet as the CNN Encoder + Decoder</li> <li>▪ 12-layer ViT as the Transformer Encoder with ImageNet pretrained weights</li> <li>▪ A hybrid loss function consisting of pixel-wise cross entropy loss and dice loss</li> <li>▪ Pre-train the transformer blocks using Masked Autoencoder (He et al., 2022)</li> <li>▪ Code reference: 3D TransUNet Model</li> </ul>
blackbean (STU-Net)	<ul style="list-style-type: none"> <li>▪ A scalable and transferable version of nnUNet</li> <li>▪ Larger input patch size: 160 × 160 × 160</li> <li>▪ Poly decay policy</li> <li>▪ Code reference: STU-NET and nnUNetV1</li> </ul>
CNMCPMI2023 (Label-wise model ensemble approach)	<ul style="list-style-type: none"> <li>▪ nnU-Net and Swin UNETR CNN + ViT</li> <li>▪ Outputs of these networks are then subjected to a non-linear function</li> <li>▪ Processed outputs are combined through model ensembling to create ensembled predictions</li> <li>▪ Label-wise post-processing is then applied to these ensembled predictions to produce the final predictions for each label</li> </ul>

ment history and assess treatment response. Presence of BMs is often a prognostic indicator of poor outcome in patients with metastatic disease, significantly changing treatment options and impacting patient survival (Jekel et al., 2022a; Chen et al., 2023b; Ottesen et al., 2023). The 2023 BraTS-METS challenge has significantly driven forward

the development of algorithms designed to manage the complex task of BMs segmentation. These algorithms provide clinicians with better tools to measure tumor volumes accurately, which is crucial for both treatment planning and patient outcomes. The varying performance among the participating teams underlines the inherent complexity

Table 6: Teams' Dice scores, reported as mean  $\pm$  standard deviation (median), and ranking based on individual tumor entities.

Team Name	ET		TC		WT	
	Dice score	Rank	Dice score	Rank	Dice score	Rank
NVAUTO	0.60 $\pm$ 0.24 (0.58)	1	0.65 $\pm$ 0.25 (0.60)	1	0.62 $\pm$ 0.24 (0.61)	1
SY	0.57 $\pm$ 0.28 (0.57)	2	0.62 $\pm$ 0.29 (0.64)	2	0.60 $\pm$ 0.29 (0.61)	2
blackbean	0.57 $\pm$ 0.26 (0.58)	2	0.61 $\pm$ 0.28 (0.58)	3	0.57 $\pm$ 0.28 (0.57)	4
CNMCPMI2023	0.55 $\pm$ 0.28 (0.64)	4	0.60 $\pm$ 0.30 (0.69)	4	0.58 $\pm$ 0.29 (0.64)	3
isahajmistry	0.49 $\pm$ 0.29 (0.44)	5	0.53 $\pm$ 0.29 (0.49)	5	0.48 $\pm$ 0.27 (0.43)	5
DeepRadOnc	0.39 $\pm$ 0.31 (0.39)	6	0.43 $\pm$ 0.36 (0.43)	6	0.40 $\pm$ 0.31 (0.41)	7
MIASINTEF	0.39 $\pm$ 0.29 (0.39)	6	0.43 $\pm$ 0.31 (0.44)	6	0.43 $\pm$ 0.32 (0.43)	6

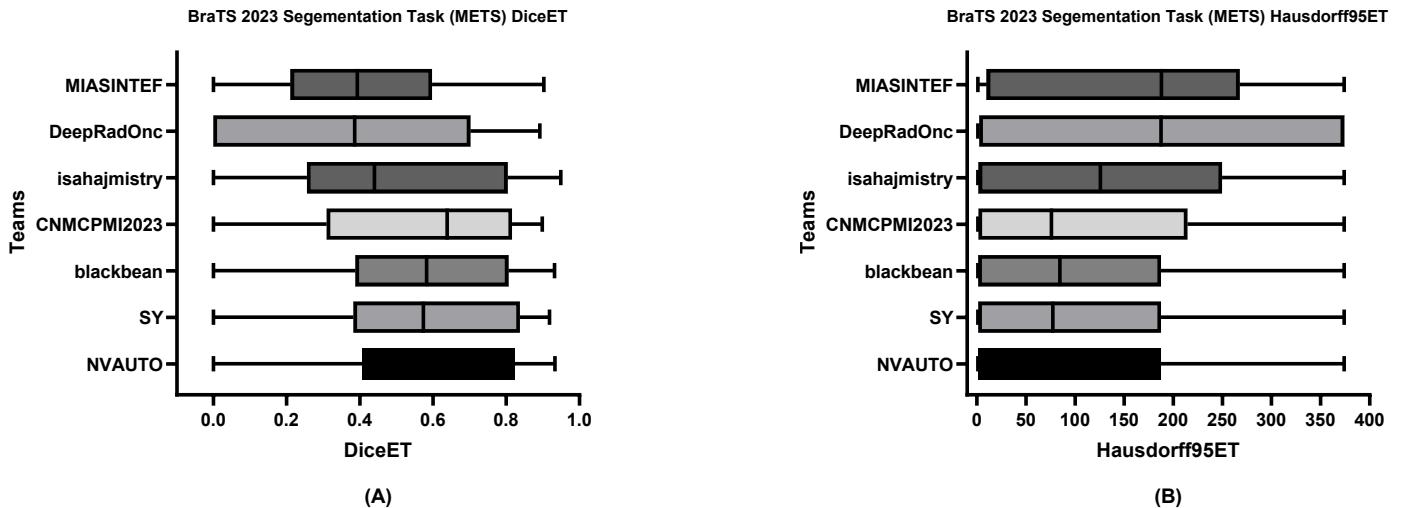


Figure 6: BraTS-METS 2023 boxplots of enhancing tumor Dice scores (A) and 95% Hausdorff distance (HD95) (B) for all participating teams on the BraTS 2023 test set.

of tumor segmentation in diverse datasets. This diversity in results particularly highlights the difficulty algorithms face in consistently identifying and accurately segmenting small metastases, which remain a significant hurdle in the literature, clinical practice, and for BraTS-METs challenge participants. The assessment metric utilized in BraTS-METs 2023 challenge penalizes for false negatives and false positives, which provides overall low Dice coefficients but provides a metric that optimizes for selection of algorithms that will be easily translated into diverse clinical practices. The performance trends observed in the challenge demonstrate that while some progress has been made, the precise detection of small metastases continues to be the princi-

pal challenge, limiting the overall effectiveness of current models. Enhancing the sensitivity and specificity of these models for small lesion detection is crucial, as this would lead to significant improvements in diagnostic accuracy and clinical outcomes. Improving sensitivity of small metastases will likely require both larger sample sizes and novel network architectures or loss functions that focus on lesionwise detection as currently employed loss functions are optimized towards voxelwise performance.

While multiple algorithms have shown promise in accurately segmenting BMs with high Dice scores (Dikici et al., 2020, 2022; Charron et al., 2018; Bousabarah et al., 2020), a critical limitation remains in their ability to detect very

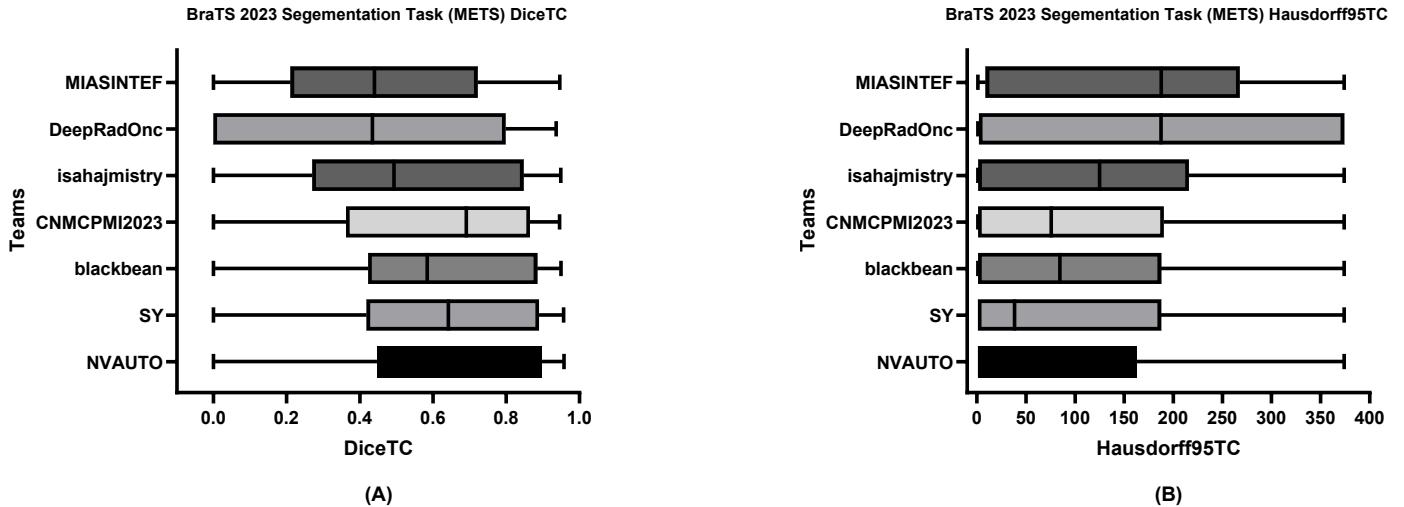


Figure 7: BraTS-METS 2023 boxplots of tumor core Dice scores (A) and 95% Hausdorff distance (HD95) (B) for all participating teams on the BraTS 2023 test set.

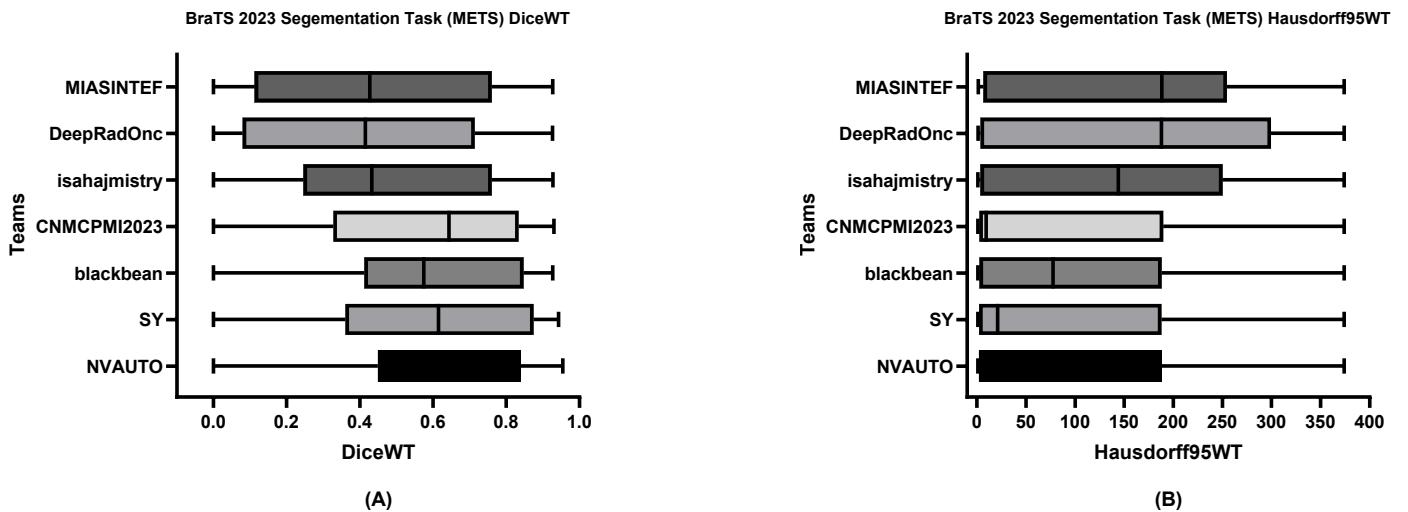


Figure 8: BraTS-METS 2023 boxplots of whole tumor Dice scores (A) and 95% Hausdorff distance (HD95) (B) for all participating teams on the BraTS 2023 test set.

small lesions, i.e., under 5 mm in size. Accurately identifying and quantifying every lesion, regardless of size, is paramount for effective therapeutic planning and prognosis assessment. Fairchild et al. (2024) retrospectively investigated BMs that were missed on initial MRIs, despite meeting diagnostic criteria, but became detected upon subsequent imaging in patients undergoing repeat SRS courses (Fairchild et al., 2024). The radiographic evidence of these metastases could often be spotted in earlier scans, suggesting potential for improved early detection and treatment planning. This issue is particularly pronounced for lesions under 3 mm, which may go untreated initially, only to become apparent on future imaging (Fairchild et al., 2023).

The heterogeneity in the appearance of BMs—ranging from multiple small lesions to solitary large lesions with varying degrees of edema—presents unique challenges in their detection and management. Our review of the challenge

outcomes shows that Team NVAUTO achieved the highest scores, with a mean lesion-wise Dice score of 0.60 to 0.65 across different tumor entities. While these results place them at the forefront, the scores also highlight that there is considerable potential for further advancements. The close performance of teams like SY and blackbean illustrates the competitive nature of the field and emphasizes the need for ongoing improvements in precision, especially for smaller and more challenging lesions.

It is essential to highlight how various models developed for the 2023 BraTS-METS challenge handled the segmentation of these critical, small lesions. Our analysis of model performance across different lesion sizes revealed significant variations in how these models managed lesion detection and characterization. For instance, NVAUTO exhibited exceptional performance across all lesion sizes, particularly with smaller lesions, surpassing the overall per-

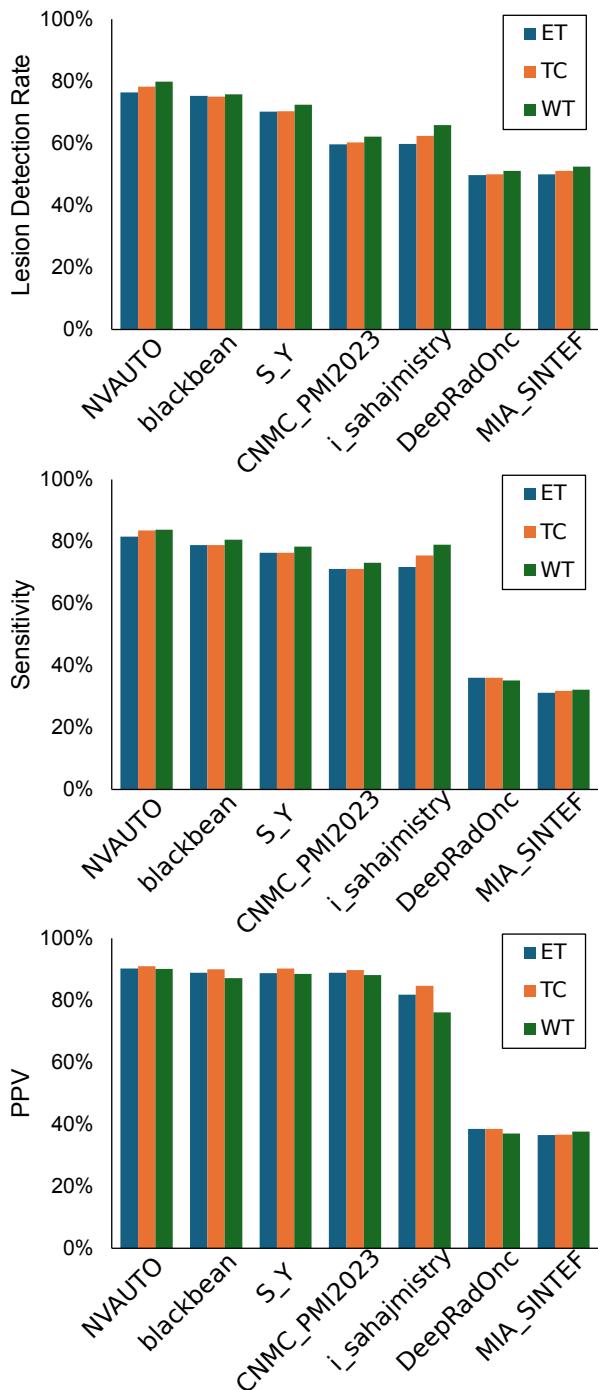


Figure 9: Performance metrics across tumor entities—whole tumor (WT), tumor core (TC), and enhancing tumor (ET).

formance of many other models in the challenge. These model performance findings underscore the necessity for continuous improvement in the algorithms’ sensitivity to tumor size variations, which is crucial for ensuring that all lesions, particularly the smaller and potentially more elusive ones, are accurately identified and appropriately managed in clinical settings.

In the realm of targeted therapies, such as radiation, precision in lesion segmentation directly influences treat-

ment efficacy, as determining lesion sizes influences SRS dose. For example, lesions up to 20 mm may receive up to 24 Gy, which is adjusted based on the lesion’s diameter to prevent severe neurotoxicity (Shaw et al., 2000). Misidentifying or overlooking even a single small lesion can lead to inadequate treatment coverage, potentially resulting in suboptimal patient outcomes and increased recurrence rates (Kaal et al., 2005; Zindler et al., 2014). This underscores the necessity for advancements in diagnostic imaging techniques and highlights the critical role of machine learning technologies in achieving high precision in BMs detection and segmentation. In turn, these algorithms have the potential to significantly impact treatment response assessments and improve workflow efficiencies in clinical practice.

Accurate detection and precise quantification of lesion volumes are critical for determining patient prognosis. Prior research has shown that the GTV of metastatic disease within the brain significantly impacts patient survival, particularly when deciding between equivalent treatment options such as surgery and radiotherapy (Routman et al., 2018; Krist et al., 2022). This precise volume measurement helps clinicians choose the most appropriate therapeutic approach, ensuring that treatments like SRS or invasive surgical interventions are tailored to the patient’s specific disease burden.

The ability to assess the GTV of BMs at diagnosis is crucial for patient outcomes. Accurately tracking changes in lesion volumes and perilesional edema over time is essential for informed decision-making in the post-treatment setting (Jalalifar et al., 2023). Treatments for brain metastatic disease utilize targeted approaches such as SRS, hypofractionated stereotactic radiation therapy (HFSRT), and hippocampal avoidance whole brain radiotherapy with less common use of whole brain radiation therapy due to neurotoxicity concerns. These techniques are particularly beneficial for patients with multiple metastases—even over 50—and rely heavily on precise volumetric localization of each metastasis (Simon et al., 2022). Unlike WBRT, which uses a 2D plan and does not require detailed localization, SRS and HFSRT involve complex 3D planning to accurately target each lesion. Furthermore, the dynamic nature of these metastases—with some increasing in size transiently before decreasing or resolving, and others possibly representing radiation necrosis or recurrence—underscores the necessity for reliable monitoring of metastasis sizes in relation to treatment timing (Wang et al., 2023a). This ongoing surveillance of the contrast enhancing component and peri-tumoral edema is vital to differentiate between active disease and treatment effects, thereby guiding the adjustment of therapeutic strategies (Kaur et al., 2023; Jekel et al., 2022a).

A significant challenge in creating large open science datasets involves safeguarding patient privacy and securing

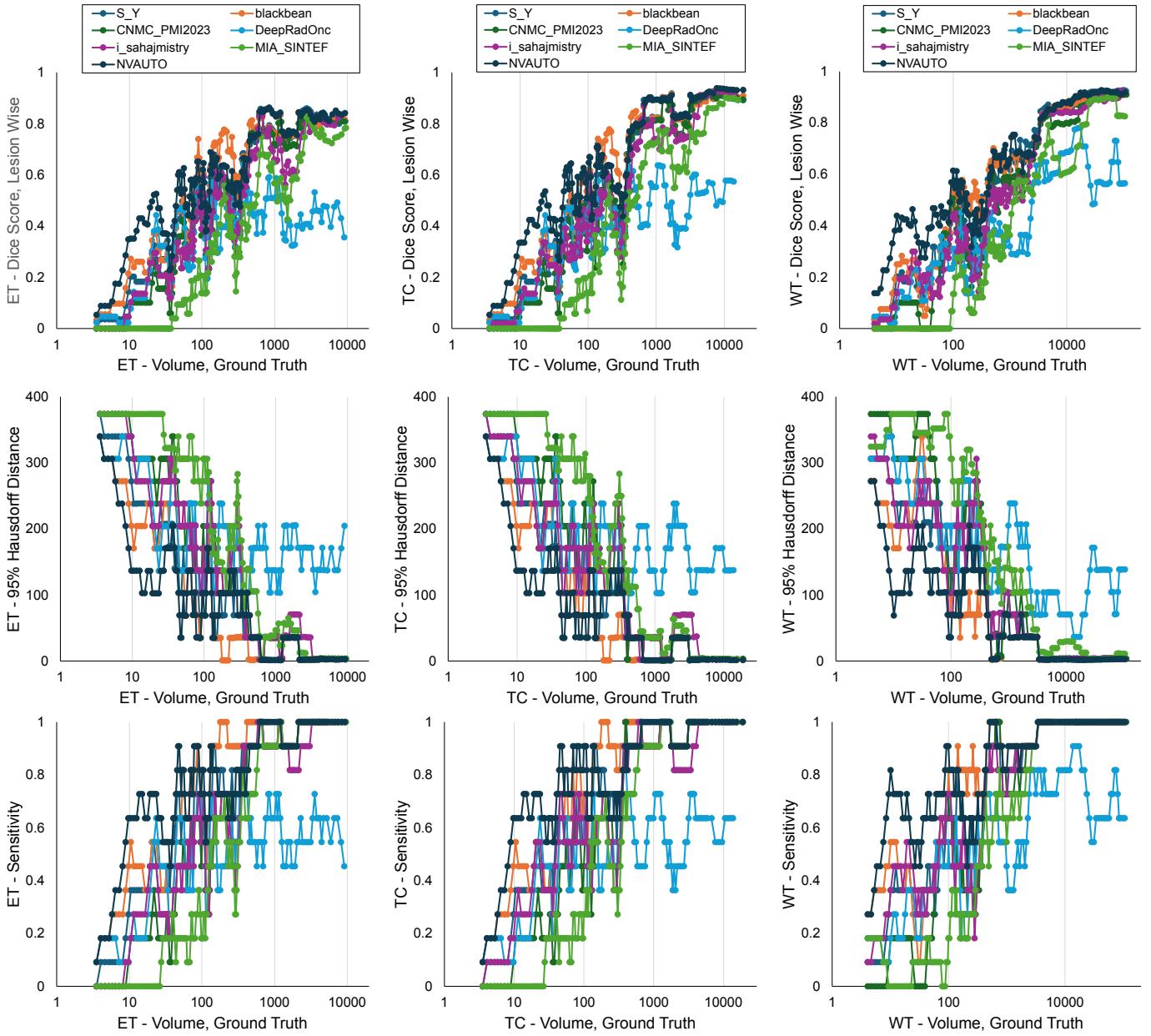


Figure 10: BraTS-METS 2023 plot of cumulative average of (A) Dice scores, (B) 95% Hausdorff distance (HD95), and (C) lesion detection rate as a function of increasing lesion volume.

sensitive data (Vahdati et al., 2024; Shaw et al., 2024; Wang et al., 2024; Gichoya et al., 2023; Davis et al., 2024). This can be addressed by establishing robust security measures, such as data de-identification using skull and face stripping from the MRI scan to remove facial features. Moreover, fostering a culture of sharing and collaboration is essential for the broad applicability of these algorithms across different institutions. It is vital to balance promoting open science with maintaining patient safety, as this balance will drive future advancements in medical image analysis. This focus on open science not only broadens access to data but also introduces challenges in data handling and annotation, particularly for complex cases like BMs.

In the 2023 inaugural BraTS-METS challenge, a sig-

nificant hurdle was the preparation of BMs datasets with expert-approved lesion annotations. Unlike other brain tumors such as glioblastomas or meningiomas, BMs display significant phenotypic variability and are often characterized by the presence of multiple synchronous lesions. This variability and multiplicity greatly complicate the annotation process, extending the time required from a few minutes to several hours depending on the number and complexity of lesions.

To address this, we introduced an innovative educational approach to annotation that not only facilitates the development of high-quality annotated datasets but also serves as a learning platform for annotators. This strategy involves a comprehensive educational series on BM imag-

ing, basic MRI physics, and the principles of open science. This approach emphasizes deliberate learning (Mitchell and Boyer, 2020), where student annotators engage deeply with the material through practical experience, reinforced by weekly hands-on sessions with experts in brain tumor imaging and a structured curriculum. This method not only accelerates the learning curve but also ingrains a thorough comprehension of diverse BM presentations, turning the annotation process into a valuable educational experience and creating a rich training resource for future professionals. Additionally, the curriculum includes detailed discussions on various brain abnormalities such as microvascular white matter damage, microbleeds, and different stages of hemorrhage, further enriching their understanding and capabilities in annotating complex imaging datasets.

While our approach faced challenges due to the heterogeneity of the contributed datasets, this diversity is reflective of real-world clinical environments where algorithms must perform effectively across a wide range of data variations. Many cases were excluded from the analysis due to resection cavities, post-treatment changes, or the absence of brain parenchymal metastases. Inadequate skull stripping sometimes led to the inadvertent removal of metastases or failure to detect them, complicating accurate data interpretation. Furthermore, skull stripping can make it difficult to describe and differentiate dural-based lesions, such as metastases and meningiomas, and limits the evaluation of osseous metastases to the calvarium.

Another source of heterogeneity was due to differences in data acquisition, patient motion, protocols, slice thickness, and contrast injection timing that can lead to misregistration of images on different sequences. Particularly, the impact of slice thickness on lesion detectability is crucial, especially when targeting subcentimeter metastases. For example, the RANO high grade glioma criteria specify lesion visibility on two contiguous 5 mm thick slices, underscoring the importance of image resolution (Wen et al., 2023). During our manual segmentation processes, challenges arose when matching sequences acquired with varying 2D and 3D techniques, highlighting disparities in slice thickness and voxel sizes. In some instances, the co-registration of images appeared misaligned, potentially affecting the precision of segmentations. To address some of these issues, all images were standardized by registering them to the common SRI24 atlas (Rohlfing et al., 2010), promoting greater uniformity and adherence to the consensus brain tumor imaging protocol. This not only helped to mitigate the variations introduced by different imaging protocols but also enhanced the general applicability and effectiveness of the developed algorithms. These limitations contribute to the heterogeneity of data, which can have both positive and negative implications. While it can pose challenges for developing a uniform segmentation algorithm, it can

also provide a diverse range of data that can benefit and generalize algorithm development.

While standardization of brain tumor imaging protocols (BTIP) have been proposed and are increasingly used in clinical trials resulting improved standardization of image acquisition, there is still a significant variability in imaging protocols among different imaging practices (Ellingson et al., 2021, 2015; Kaufmann et al., 2020). Increased implementation of standardized imaging protocols ensures consistency in the acquisition and interpretation of neuro-oncological images, which is crucial for comparing outcomes across studies and improving the reliability of lesion measurement across different institutions.

The complexity of annotating ground truth data for BMs represents yet another challenge in this year's BraTS-METS challenge, largely due to the typically small size of BMs and their frequent occurrence in large numbers within a single scan. Annotator fatigue is a notable concern, as the meticulous nature of the task can lead to errors or oversight. Throughout the annotation process, numerous instances necessitated segmentation revisions, as exemplified by the initial work done on the Yale BM dataset by a medical student, which later required refinement by experienced neuroradiologists (Kaur et al., 2023; Cassinelli Petersen et al., 2022; Jekel et al., 2022a; Ramakrishnan et al., 2023). The need for such revisions became particularly apparent when the dataset, along with its segmentations, was integrated into the BraTS challenge and adapted to a new atlas. This process often revealed previously unnoticed small lesions or inaccuracies in the depiction of necrotic tumor portions and peritumoral edema on FLAIR images. These experiences showcase the imperative of a robust ground truth (i.e. reference standard) approach that incorporates humans in the loop refinements and utilizes consensus techniques like STAPLE to ensure the highest data integrity (Warfield et al., 2004). The iterative nature of these annotations underscores the need for multiple rounds of review to ensure accuracy and the importance of standardizing annotation practices to facilitate more efficient data usage. To foster continual improvement and address any discrepancies, we encourage participants to engage actively with the challenge organizers, who are prepared to update and refine the segmentation data as necessary to maintain the integrity and utility of the dataset.

## 7. Conclusion

In the inaugural 2023 BraTS-METS challenge, we have addressed both technical and practical challenges in the establishment of datasets, high quality reference standard annotations, and assessment metrics for the development and application of machine learning algorithms for BM segmentation by challenge participants. The challenge has

highlighted the critical need for algorithms capable of detecting even the smallest lesions, which are often overlooked due to human error or obscured by the limitations of imaging data. This task is complicated by the necessity of balancing the high sensitivity required for detection with the need to minimize false positives that can disrupt clinical workflows. The development of refined segmentation algorithms that effectively balance sensitivity with specificity is therefore essential. Utilizing multi-institutional datasets, the BraTS-METS challenge has been instrumental in advancing these developments, pushing forward the creation of models that are robust and adaptable across varied clinical environments. This approach optimizes the precision of these algorithms and potentiates their practical applicability, ensuring they can meet the nuanced demands of real-world medical practice. As we continue to refine these technologies, our goal remains to enhance the accuracy of diagnoses and treatment planning, ultimately improving patient management and outcomes in the challenging arena of brain metastasis treatment.

## Acknowledgments

The success of any challenge in the medical domain depends upon the quality of well-annotated multi-institutional datasets. We are grateful to all the data contributors, annotators, and approvers for their time and efforts. We are grateful to the institutions that contributed directly and indirectly to resources for the development of the databases. We are also grateful to individual companies that assisted in the development of datasets, such as Visage Imaging in the development of the Yale BM dataset.

S. Bakas and U. Baid conducted part of the work reported in this manuscript at their current affiliation, as well as while they were affiliated with the Center for Artificial Intelligence and Data Science for Integrated Diagnostics (AI2D) and the Center for Biomedical Image Computing and Analytics (CBICA), Perelman School of Medicine at the University of Pennsylvania, Philadelphia.

M. Aboian conducted part of the work reported in this manuscript at her current affiliation, as well as while she was affiliated with Yale University School of Medicine, New Haven, CT.

We thank Victoria Ramirez (Department of Radiology, Children's Hospital of Philadelphia) for her efforts in reviewing the manuscript.

We thank Ananya Purwar for her technical support in editing the LaTeX formatting for this work.

## Funding

Research reported in this publication was partly supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH), under award numbers U01CA242871, NIH/NCI R21CA259964. The research was supported by Yale Department of Radiology and by Children's Hospital of Philadelphia (CHOP) Department of Radiology. The content of this publication is the sole responsibility of the authors and does not represent the official views of the NIH.

## Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

## Conflicts of Interest

No conflicts of interest to disclose.

## Data availability

The data provided for the challenge is available on the Challenge Page Link. All the analysis will be shared via BOX on request.

## References

Khaled Bousabarah, Maximilian Ruge, Julia-Sarita Brand, Mauritius Hoevels, Daniel Rueß, Jan Borggrefe, Nils Große Hokamp, Veerle Visser-Vandewalle, David Maintz, Harald Treuer, et al. Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data. *Radiation Oncology*, 15:1–9, 2020.

Josef A Buchner, Jan C Peeken, Lucas Etzel, Ivan Ezhov, Michael Mayinger, Sebastian M Christ, Thomas B Brunner, Andrea Wittig, Bjoern H Menze, Claus Zimmer, et al. Identifying core mri sequences for reliable automatic brain metastasis segmentation. *Radiotherapy and Oncology*, 188:109901, 2023.

Gabriel Cassinelli Petersen, Khaled Bousabarah, Tej Verma, Marc von Reppert, Leon Jekel, Ayyuce Gordem, Benjamin Jang, Sara Merkaj, Sandra Abi Fadel, Randy Owens, et al. Real-time pacs-integrated longitudinal brain metastasis tracking tool provides comprehensive assessment of treatment response to radiosurgery. *Neuro-Oncology Advances*, 4(1):vdac116, 2022.

Odelin Charron, Alex Lallement, Delphine Jarnet, Vincent Noblet, Jean-Baptiste Clavier, and Philippe Meyer. Auto-

- matic detection and segmentation of brain metastases on multimodal mr images with a deep convolutional neural network. *Computers in biology and medicine*, 95:43–54, 2018.
- Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. 3d transunet: Advancing medical image segmentation through vision transformers. *arXiv preprint arXiv:2310.07781*, 2023a.
- Mingming Chen, Yujie Guo, Pengcheng Wang, Qi Chen, Lu Bai, Shaobin Wang, Ya Su, Lizhen Wang, and Guanzhong Gong. An effective approach to improve the automatic segmentation and classification accuracy of brain metastasis by combining multi-phase delay enhanced mr images. *Journal of Digital Imaging*, 36(4):1782–1793, 2023b.
- Victor Eric Chen, Minchul Kim, Nicolas Nelson, Inkyu Kevin Kim, and Wenyin Shi. Cost-effectiveness analysis of 3 radiation treatment strategies for patients with multiple brain metastases. *Neuro-Oncology Practice*, 10(4):344–351, 2023c.
- Se Jin Cho, Leonard Sunwoo, Sung Hyun Baik, Yun Jung Bae, Byung Se Choi, and Jae Hyoung Kim. Brain metastasis detection using machine learning: a systematic review and meta-analysis. *Neuro-oncology*, 23(2):214–225, 2021.
- NP Dang, G Noid, Y Liang, JA Bovi, M Bhalla, and A Li. Automated brain metastasis detection and segmentation using deep-learning method. *International Journal of Radiation Oncology, Biology, Physics*, 114(3):e50, 2022.
- Melissa A Davis, Ona Wu, Ichiro Ikuta, John E Jordan, Michele H Johnson, and Edward Quigley. Understanding bias in artificial intelligence: A practice perspective. *American Journal of Neuroradiology*, 45(4):371–373, 2024.
- Engin Dikici, John L Ryu, Mutlu Demirer, Matthew Bigelow, Richard D White, Wayne Slone, Barbaros Selnur Erdal, and Luciano M Prevedello. Automated brain metastases detection framework for t1-weighted contrast-enhanced 3d mri. *IEEE journal of biomedical and health informatics*, 24(10):2883–2893, 2020.
- Engin Dikici, Xuan V Nguyen, Matthew Bigelow, John L Ryu, and Luciano M Prevedello. Advancing brain metastases detection in t1-weighted contrast-enhanced 3d mri using noisy student-based training. *Diagnostics*, 12(8):2023, 2022.
- Benjamin M Ellingson, Martin Bendszus, Jerrold Boxerman, Daniel Barboriak, Bradley J Erickson, Marion Smits, Sarah J Nelson, Elizabeth Gerstner, Brian Alexander, Gregory Goldmacher, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro-oncology*, 17(9):1188–1198, 2015.
- Benjamin M Ellingson, Matthew S Brown, Jerrold L Boxerman, Elizabeth R Gerstner, Timothy J Kaufmann, Patricia E Cole, Jeffrey A Bacha, David Leung, Amy Barone, Howard Colman, et al. Radiographic read paradigms and the roles of the central imaging laboratory in neuro-oncology clinical trials. *Neuro-oncology*, 23(2):189–198, 2021.
- Andrew Fairchild, Joseph K Salama, Devon Godfrey, Walter F Wiggins, Bradley G Ackerson, Taofik Oyekunle, Donna Niedzwiecki, Peter E Fecci, John P Kirkpatrick, and Scott R Floyd. Incidence and imaging characteristics of difficult to detect retrospectively identified brain metastases in patients receiving repeat courses of stereotactic radiosurgery. *Journal of Neuro-Oncology*, pages 1–9, 2024.
- Andrew T Fairchild, Joseph K Salama, Walter F Wiggins, Bradley G Ackerson, Peter E Fecci, John P Kirkpatrick, Scott R Floyd, and Devon J Godfrey. A deep learning-based computer aided detection (cad) system for difficult-to-detect brain metastases. *International Journal of Radiation Oncology\* Biology\* Physics*, 115(3):779–793, 2023.
- Florin C Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Dominik Neumann, Pragneshkumar Patel, Reddappagari Suryanarayana Vishwanath, James M Balter, Yue Cao, Sasa Grbic, et al. Contrastive self-supervised learning from 100 million medical images with optional supervision. *Journal of Medical Imaging*, 9(6):064503–064503, 2022.
- Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. Ai pitfalls and what not to do: mitigating bias in ai. *The British Journal of Radiology*, 96(1150):20230023, 2023.
- Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging*, 35(5):1153–1159, 2016.
- Endre Grøvik, Darvin Yi, Michael Iv, Elizabeth Tong, Daniel Rubin, and Greg Zaharchuk. Deep learning enables automatic detection and segmentation of brain metastases

- on multisequence mri. *Journal of Magnetic Resonance Imaging*, 51(1):175–182, 2020.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.
- Seyed Ali Jalalifar, Hany Soliman, Arjun Sahgal, and Ali Sadeghi-Naini. Automatic assessment of stereotactic radiation therapy outcome in brain metastasis using longitudinal segmentation on serial mri. *IEEE Journal of Biomedical and Health Informatics*, 2023.
- Leon Jekel, Khaled Bousabarah, MingDe Lin, Sara Merkaj, Manpreet Kaur, Arman Avesta, Sanjay Aneja, Antonio Omuro, Veronica Chiang, Björn Scheffler, et al. Nimg-02. pacs-integrated auto-segmentation workflow for brain metastases using nnu-net. *Neuro-oncology*, 24(Supplement\_7):vii162–vii162, 2022a.
- Leon Jekel, Waverly R Brim, Marc von Reppert, Lawrence Staib, Gabriel Cassinelli Petersen, Sara Merkaj, Harry Subramanian, Tal Zeevi, Seyedmehdi Payabvash, Khaled Bousabarah, et al. Machine learning applications for differentiation of glioma from brain metastasis—a systematic review. *Cancers*, 14(6):1369, 2022b.
- Hana Jeong, Ji Eun Park, NakYoung Kim, Shin-Kyo Yoon, and Ho Sung Kim. Deep learning-based detection and quantification of brain metastases on black-blood imaging can provide treatment suggestions: a clinical cohort study. *European Radiology*, 34(3):2062–2071, 2024.
- Krishna Juluru, Eliot Siegel, and Jan Mazura. Identification from mri with face-recognition software. *The New England Journal of Medicine*, 382(5):489–490, 2020.
- Evert CA Kaal, Charles GJH Niël, and Charles J Vecht. Therapeutic management of brain metastasis. *The Lancet Neurology*, 4(5):289–298, 2005.
- Hemalatha Kanakarajan, Wouter De Baene, Patrick Hanssens, and Margriet Sitskoorn. Fully automated brain metastases segmentation using t1-weighted contrast-enhanced mr images before and after stereotactic radiosurgery. *medRxiv*, pages 2023–07, 2023.
- Timothy J Kaufmann, Marion Smits, Jerrold Boxerman, Raymond Huang, Daniel P Barboriak, Michael Weller, Caroline Chung, Christina Tsien, Paul D Brown, Lalitha Shankar, et al. Consensus recommendations for a standardized brain tumor imaging protocol for clinical trials in brain metastases. *Neuro-oncology*, 22(6):757–772, 2020.
- Manpreet Kaur, Gabriel Cassinelli Petersen, Leon Jekel, Marc von Reppert, Sunitha Varghese, Irene Dixe de Oliveira Santo, Arman Avesta, Sanjay Aneja, Antonio Omuro, Veronica Chiang, et al. Pacs-integrated tools for peritumoral edema volumetrics provide additional information to rano-bm-based assessment of lung cancer brain metastases after stereotactic radiotherapy: A pilot study. *Cancers*, 15(19):4822, 2023.
- Linda C Kelahan, Donald Kim, Moataz Soliman, Ryan J Avery, Hatice Savas, Rishi Agrawal, Michael Magnetta, Benjamin P Liu, and Yuri S Velichko. Role of hepatic metastatic lesion size on inter-reader reproducibility of ct-based radiomics features. *European radiology*, 32(6):4025–4033, 2022.
- David T Krist, Anant Naik, Charee M Thompson, Susanna S Kwok, Mika Janbahan, William C Olivero, and Wael Hassaneen. Management of brain metastasis. surgical resection versus stereotactic radiotherapy: a meta-analysis. *Neuro-Oncology Advances*, 4(1):vdac033, 2022.
- E Le Rhun, Matthias Guckenberger, Marion Smits, Reinhard Dummer, Thomas Bachelot, Felix Sahm, Norbert Galldiks, Evandro de Azambuja, Anna Sophie Berghoff, Philippe Metellus, et al. Eano-esmo clinical practice guidelines for diagnosis, treatment and follow-up of patients with brain metastasis from solid tumours. *Annals of Oncology*, 32(11):1332–1347, 2021.
- Andrea Liew, Chun Cheng Lee, Valarmathy Subramaniam, Boon Leong Lan, and Maxine Tan. Gradual self-training via confidence and volume based domain adaptation for multi dataset deep learning-based brain metastases detection using nonlocal networks on mri images. *Journal of Magnetic Resonance Imaging*, 57(6):1728–1740, 2023.
- Nancy U Lin, Eudocia Q Lee, Hidefumi Aoyama, Igor J Barani, Daniel P Barboriak, Brigitta G Baumert, Martin Bendszus, Paul D Brown, D Ross Camidge, Susan M Chang, et al. Response assessment criteria for brain metastases: proposal from the rano group. *The lancet oncology*, 16(6):e270–e278, 2015.
- Oskar Maier, Bjoern H Menze, Janina Von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen,

- et al. Isles 2015-a public evaluation benchmark for is-chemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017.
- Honglan Mi, Mingyuan Yuan, Shiteng Suo, Jiejun Cheng, Suqin Li, Shaofeng Duan, and Qing Lu. Impact of different scanners and acquisition parameters on robustness of mr radiomics features based on women's cervix. *Scientific reports*, 10(1):20407, 2020.
- Giuseppe Minniti, Enrico Clarke, Gaetano Lanzetta, Mattia Falchetto Osti, Guido Trasimeni, Alessandro Bozzao, Andrea Romano, and Riccardo Maurizi Enrici. Stereotactic radiosurgery for brain metastases: analysis of outcome and risk of brain radionecrosis. *Radiation oncology*, 6: 1–9, 2011.
- Sally A Mitchell and Tanna J Boyer. Deliberate practice in medical simulation. 2020.
- Reabal Najjar. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17):2760, 2023.
- Lakshmi Nayak, Eudocia Quant Lee, and Patrick Y Wen. Epidemiology of brain metastases. *Current oncology reports*, 14:48–54, 2012.
- Beatriz Ocaña-Tienda, Julián Pérez-Beteta, José D Villanueva-García, José A Romero-Rosales, David Molina-García, Yannick Suter, Beatriz Asenjo, David Albillo, Ana Ortiz de Mendivil, Luis A Pérez-Romasanta, et al. A comprehensive dataset of annotated brain metastasis mr images with clinical and radiomic data. *Scientific data*, 10(1):208, 2023.
- Eric Oermann, Katherine Link, Zane Schnurman, Chris Liu, Young Joon Fred Kwon, Lavender Yao Jiang, Mustafa Nasir-Moin, Sean Neifert, Juan Alzate, Kenneth Bernstein, et al. Longitudinal deep neural networks for assessing metastatic brain cancer on a massive open benchmark. 2023.
- Jon André Ottesen, Darvin Yi, Elizabeth Tong, Michael Iv, Anna Latysheva, Cathrine Saxhaug, Kari Dolven Jacobsen, Åslaug Helland, Kyrre Eeg Emblem, Daniel L Rubin, et al. 2.5 d and 3d segmentation of brain metastases with deep learning on multinational mri data. *Frontiers in Neuroinformatics*, 16:1056068, 2023.
- Sarthak Pati, Ashish Singh, Saima Rathore, Aimilia Gastounioti, Mark Bergman, Phuc Ngo, Sung Min Ha, Dimitrios Bounias, James Minock, Grayson Murphy, et al. The cancer imaging phenomics toolkit (captk): technical overview. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II* 5, pages 380–394. Springer, 2020.
- Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II 5, pages 380–394. Springer, 2020.
- Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah J Sheller, Patrick Foley, G Anthony Reina, Siddhesh Thakur, Chiharu Sako, Michel Bilello, Christos Davatzikos, et al. The federated tumor segmentation (fets) tool: an open-source solution to further solid tumor research. *Physics in Medicine & Biology*, 67(20):204002, 2022.
- Alan K Percy, Lila R Elveback, Haruo Okazaki, and Leonard T Kurland. Neoplasms of the central nervous system: epidemiologic considerations. *Neurology*, 22(1): 40–40, 1972.
- Irada Pflüger, Tassilo Wald, Fabian Isensee, Marianne Schell, Hagen Meredig, Kai Schlamp, Denise Bernhardt, Gianluca Brugnara, Claus Peter Heußel, Juergen Debus, et al. Automated detection and quantification of brain metastases on clinical mri data using artificial neural networks. *Neuro-oncology advances*, 4(1):vdac138, 2022.
- Luís Pinto-Coelho. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering*, 10(12):1435, 2023.
- JB Posner. Intracranial metastases from systemic cancer. *Adv. Neurol.*, 19:579–592, 1978.
- Jack M Qian, Amit Mahajan, James B Yu, A John Tsioris, Sarah B Goldberg, Harriet M Kluger, and Veronica LS Chiang. Comparing available criteria for measuring brain metastasis response to immunotherapy. *Journal of Neuro-Oncology*, 132:479–485, 2017.
- Divya Ramakrishnan, Leon Jekel, Saahil Chadha, Anastasia Janas, Harrison Moy, Nazanin Maleki, Matthew Sala, Manpreet Kaur, Gabriel Cassinelli Petersen, Sara Merkaj, et al. A large open access dataset of brain metastasis 3d segmentations with clinical and imaging feature information. *ArXiv*, 2023.
- Saima Rathore, Spyridon Bakas, Sarthak Pati, Hamed Akbari, Ratheesh Kalarot, Patmaa Sridharan, Martin Rozycki, Mark Bergman, Birkan Tunc, Ragini Verma, et al. Brain cancer imaging phenomics toolkit (brain-captk): an interactive platform for quantitative analysis of glioblastoma. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pages 133–145. Springer, 2018.

- Torsten Rohlffing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31(5):798–819, 2010.
- David M Routman, Shelly X Bian, Kevin Diao, Jonathan L Liu, Cheng Yu, Jason Ye, Gabriel Zada, and Eric L Chang. The growing importance of lesion volume as a prognostic factor in patients with multiple brain metastases treated with stereotactic radiosurgery. *Cancer medicine*, 7(3):757–764, 2018.
- Jeffrey D Rudie, Rachit Saluja, David A Weiss, Pierre Nedelec, Evan Calabrese, John B Colby, Benjamin Laguna, John Mongan, Steve Braunstein, Christopher P Hess, et al. The university of california san francisco, brain metastases stereotactic radiosurgery (ucsf-bmsr) mri dataset. *Radiology: Artificial Intelligence*, page e230126, 2024.
- Zane Schnurman, Elad Mashiach, Katherine E Link, Bernadine Donahue, Erik Sulman, Joshua Silverman, John G Golfinos, Eric Karl Oermann, and Douglas Kondziolka. Causes of death in patients with brain metastases. *Neurosurgery*, pages 10–1227, 2022.
- Christopher G Schwarz, Walter K Kremers, Terry M Therneau, Richard R Sharp, Jeffrey L Gunter, Prashanthi Vemuri, Arvin Arani, Anthony J Spychalla, Kejal Kantarci, David S Knopman, et al. Identification of anonymous mri research participants with face-recognition software. *New England Journal of Medicine*, 381(17):1684–1686, 2019.
- Edward Shaw, Charles Scott, Luis Souhami, Robert Di Napoli, Robert Kline, Jay Loeffler, and Nancy Farnan. Single dose radiosurgical treatment of recurrent previously irradiated primary brain tumors and brain metastases: final report of rtog protocol 90-05. *International Journal of Radiation Oncology\* Biology\* Physics*, 47(2):291–298, 2000.
- James Shaw, Joseph Ali, Caesar A Atuire, Phaik Yeong Cheah, Armando Guio Espa  ol, Judy Wawira Gichoya, Adrienne Hunt, Daudi Jjingo, Katherine Littler, Daniela Paolotti, et al. Research ethics and artificial intelligence for global health: perspectives from the global forum on bioethics in research. *BMC Medical Ethics*, 25(1):46, 2024.
- Mih  ly Simon, Judit Papp, Emese Csiki, and   rp  d Kov  cs. Plan quality assessment of fractionated stereotactic radiotherapy treatment plans in patients with brain metastases. *Frontiers in Oncology*, 12:846609, 2022.
- Emeline Tabouret, Olivier Chinot, Philippe Metellus, Agn  s Tallet, Patrice Viens, and Anthony Goncalves. Recent trends in epidemiology of brain metastases: an overview. *Anticancer research*, 32(11):4655–4662, 2012.
- Xiaoli Tang. The role of artificial intelligence in medical imaging research. *BJR—Open*, 2(1):20190031, 2019.
- Sanaz Vahdati, Bardia Khosravi, Elham Mahmoudi, Kuan Zhang, Pouria Rouzrokh, Shahriar Faghani, Mana Moassefi, Aylin Tahmasebi, Katherine P Andriole, Peter Chang, et al. A guideline for open-source tools to make medical imaging data ready for artificial intelligence applications: A society of imaging informatics in medicine (siim) survey. *Journal of Imaging Informatics in Medicine*, pages 1–10, 2024.
- Michael A Vogelbaum, Paul D Brown, Hans Messersmith, Priscilla K Brastianos, Stuart Burri, Dan Cahill, Ian F Dunn, Laurie E Gaspar, Na Tosh N Gatson, Vinai Gondi, et al. Treatment for brain metastases: Asco-sno-astro guideline, 2022.
- Jen-Yeu Wang, Vera Qu, Caressa Hui, Navjot Sandhu, Maria G Mendoza, Neil Panjwani, Yu-Cheng Chang, Chih-Hung Liang, Jen-Tang Lu, Lei Wang, et al. Stratified assessment of an fda-cleared deep learning algorithm for automated detection and contouring of metastatic brain tumors in stereotactic radiosurgery. *Radiation Oncology*, 18(1):61, 2023a.
- Ryan Wang, Po-Chih Kuo, Li-Ching Chen, Kenneth Patrick Seastedt, Judy Wawira Gichoya, and Leo Anthony Celi. Drop the shortcuts: image augmentation improves fairness and decreases ai detection of race and other demographics from medical images. *EBioMedicine*, 102, 2024.
- Yibin Wang, William Neil Duggar, David Michael Caballero, Toms Vengaloor Thomas, Neha Adari, Eswara Kumar Mundra, and Haifeng Wang. A brain mri dataset and baseline evaluations for tumor recurrence prediction after gamma knife radiotherapy. *Scientific Data*, 10(1):785, 2023b.
- Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.
- Patrick Y Wen, Martin van den Bent, Gilbert Youssef, Timothy F Cloughesy, Benjamin M Ellingson, Michael Weller, Evanthia Galanis, Daniel P Barboriak, John de Groot, Mark R Gilbert, et al. Rano 2.0: update to the response

assessment in neuro-oncology criteria for high-and low-grade gliomas in adults. *Journal of Clinical Oncology*, 41(33):5187–5199, 2023.

Jie Xue, Bao Wang, Yang Ming, Xuejun Liu, Zekun Jiang, Chengwei Wang, Xiyu Liu, Ligang Chen, Jianhua Qu, Shangchen Xu, et al. Deep learning-based detection and segmentation-assisted management of brain metastases. *Neuro-oncology*, 22(4):505–514, 2020.

SK Yoo, TH Kim, HJ Kim, HI Yoon, and JS Kim. Deep learning-based automatic detection and segmentation of brain metastases for stereotactic ablative radiotherapy using black-blood magnetic resonance imaging. *International Journal of Radiation Oncology, Biology, Physics*, 114(3):e558, 2022.

Youngjin Yoo, Gengyan Zhao, Andreea E Sandu, Thomas J Re, Jyoti Priya Das, Hesheng Wang, Michelle Kim, Colette Shen, Yueh Lee, Douglas Kondziolka, et al. The importance of data domain on self-supervised learning for brain metastasis detection and segmentation. In *Medical Imaging 2023: Computer-Aided Diagnosis*, volume 12465, pages 556–562. SPIE, 2023.

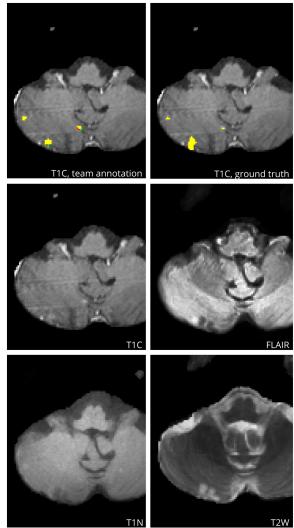
Min Zhang, Geoffrey S Young, Huai Chen, Jing Li, Lei Qin, J Ricardo McFaline-Figueroa, David A Reardon, Xinhua Cao, Xian Wu, and Xiaoyin Xu. Deep-learning detection of cancer metastases to the brain on mri. *Journal of Magnetic Resonance Imaging*, 52(4):1227–1236, 2020.

Zijian Zhou, Jeremiah W Sanders, Jason M Johnson, Maria K Gule-Monroe, Melissa M Chen, Tina M Briere, Yan Wang, Jong Bum Son, Mark D Pagel, Jing Li, et al. Computer-aided detection of brain metastases in t1-weighted mri for stereotactic radiosurgery using deep learning single-shot detectors. *Radiology*, 295(2):407–415, 2020.

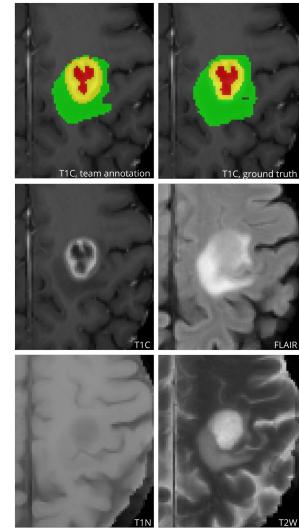
Jaap D Zindler, Ben J Slotman, and Frank J Lagerwaard. Patterns of distant brain recurrences after radiosurgery alone for newly diagnosed brain metastases: Implications for salvage therapy. *Radiotherapy and Oncology*, 112(2):212–216, 2014.

Hamidreza Ziyaee, Carlos E Cardenas, D Nana Yeboa, Jing Li, Sherise D Ferguson, Jason Johnson, Zijian Zhou, Jeremiah Sanders, Raymond Mumme, Laurence Court, et al. Automated brain metastases segmentation with a deep dive into false-positive detection. *Advances in radiation oncology*, 8(1):101085, 2023.

Case number: BraTS-MET-00137-000  
Issue: Random voxels are labelled as a NETC on the team segmentation.



Case number: BraTS-MET-00147-000  
Issue: Random voxels are labelled as a NETC on the team segmentation.



Case number: BraTS-MET-00152-000  
Issue: Random voxels are labelled as a NETC on the team segmentation.

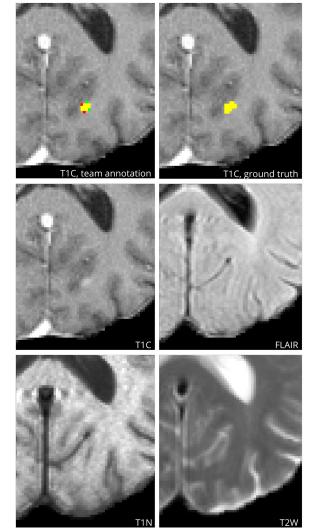
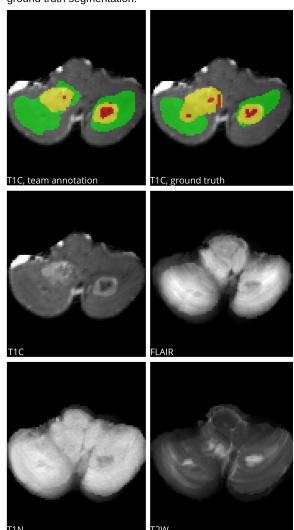
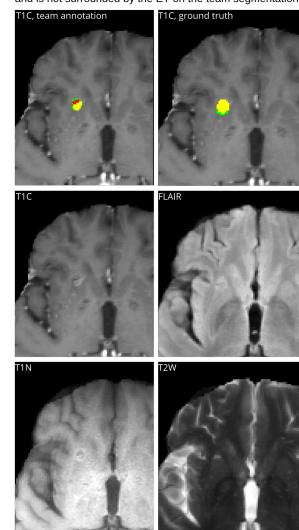


Figure 11: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

Case number: BraTS-MET-00153-000  
Issue: Random voxels are erroneously labelled as a NETC on the team segmentation. The NETC is not contained within the ET on the ground truth segmentation.



Case number: BraTS-MET-00153-000  
Issue: The NETC is not labelled on the ground truth segmentation and is not surrounded by the ET on the team segmentation.



Case number: BraTS-MET-00162-000  
Issue: Random voxel is erroneously labelled as a NETC on the team segmentation.

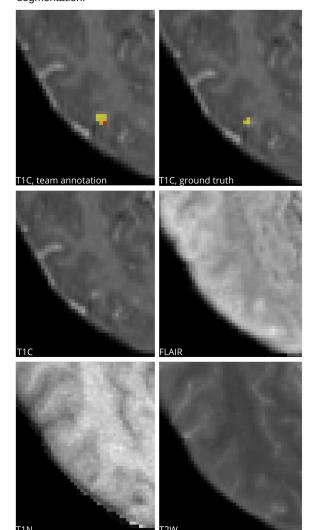


Figure 12: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

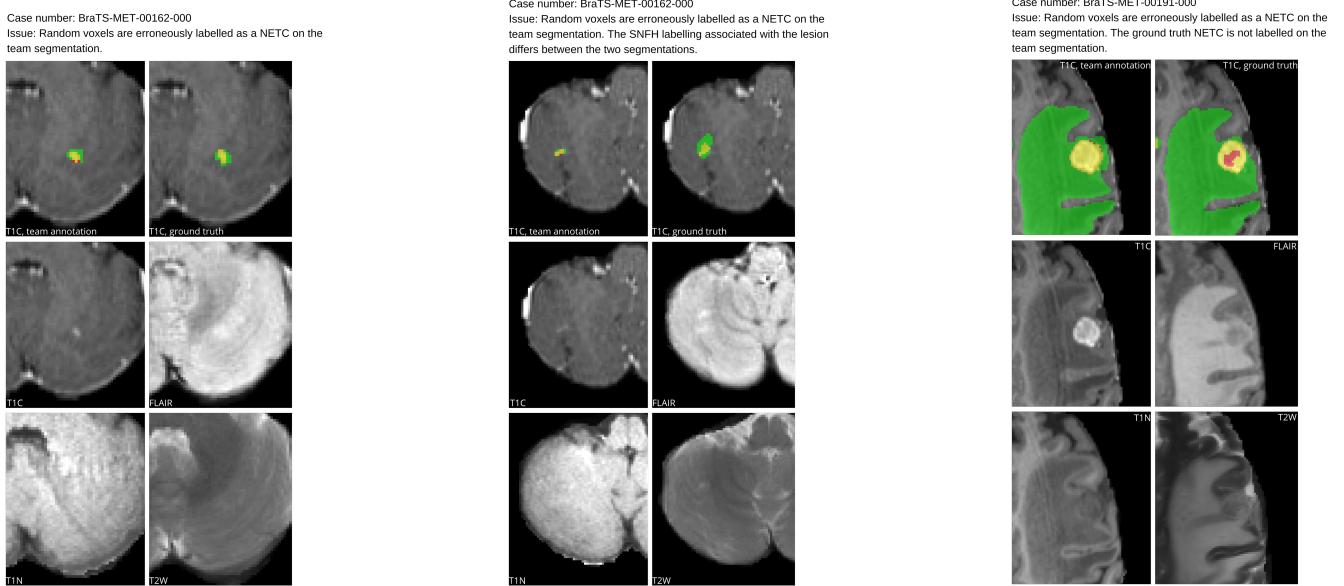


Figure 13: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

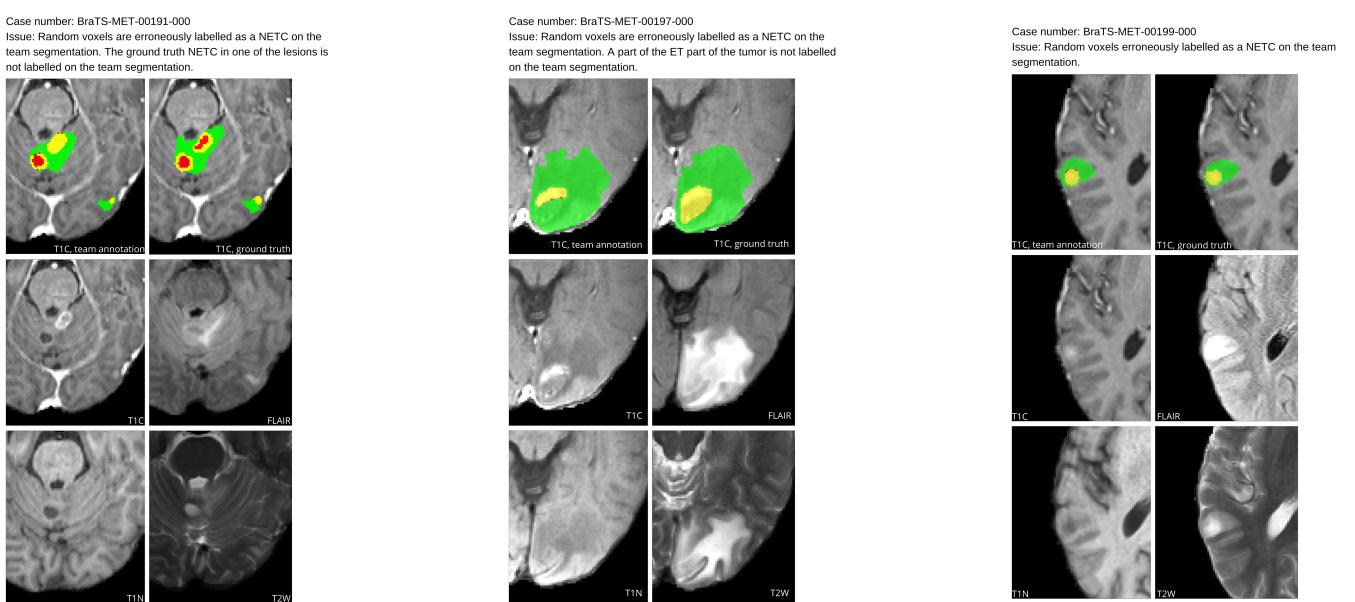
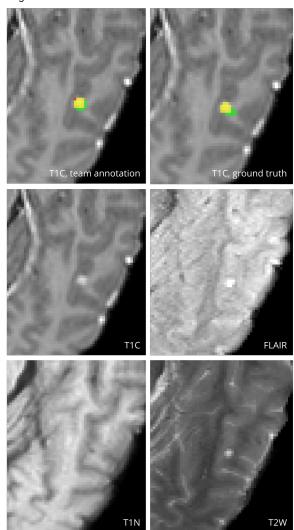
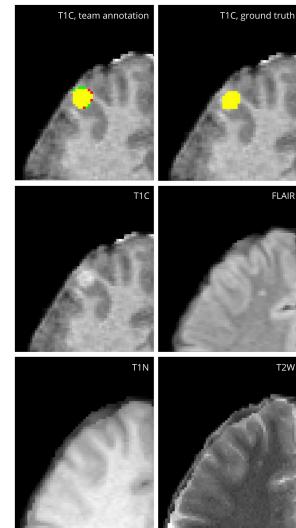


Figure 14: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

Case number: BraTS-MET-00199-000  
Issue: Random voxel erroneously labelled as a NETC on the team segmentation.



Case number: BraTS-MET-00203-000  
Issue: Random voxels erroneously labelled as a NETC on the team segmentation.



Case number: BraTS-MET-00203-000  
Issue: Random voxels erroneously labelled as a NETC on the team segmentation.

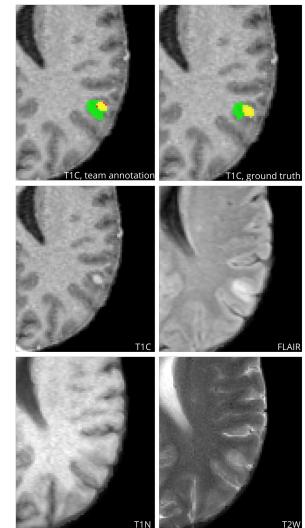
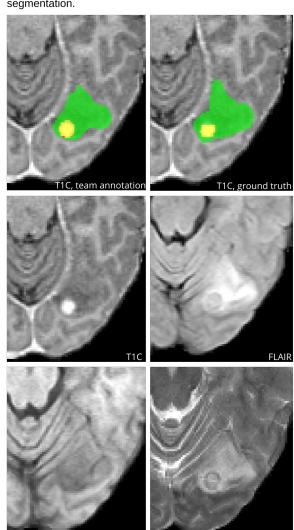
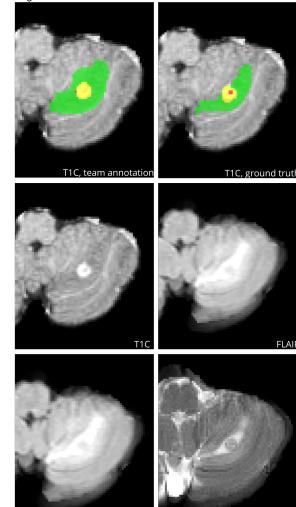


Figure 15: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

Case number: BraTS-MET-00203-000  
Issue: Random voxel erroneously labelled as a NETC on the team segmentation.



Case number: BraTS-MET-00203-000  
Issue: Random voxel erroneously labelled as a NETC on the team segmentation. The ground truth NETC is not labelled on the team segmentation.



Case number: BraTS-MET-00213-000  
Issue: Random voxels erroneously labelled as a NETC on the team segmentation.

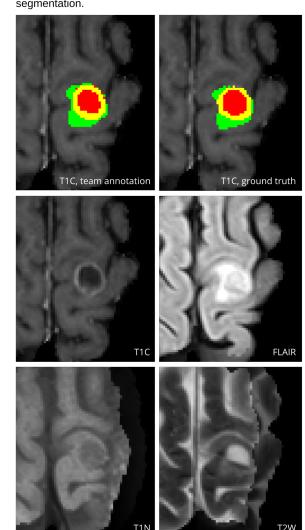


Figure 16: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

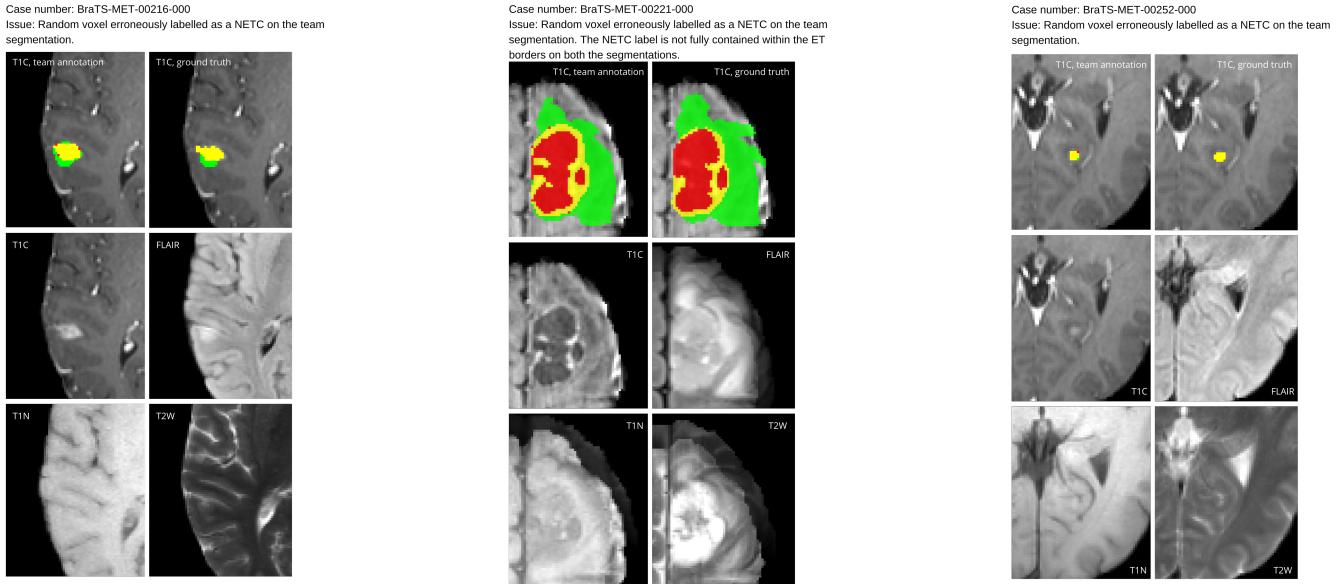


Figure 17: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

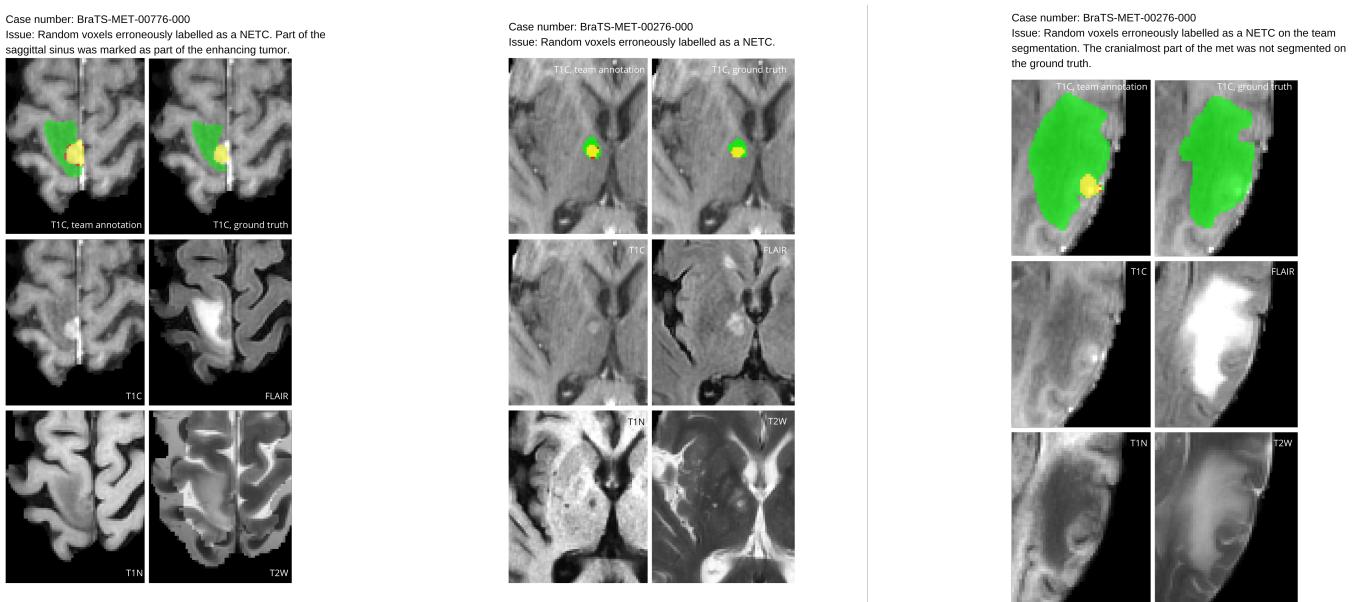


Figure 18: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

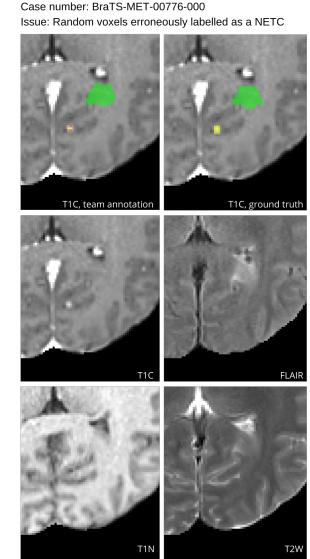
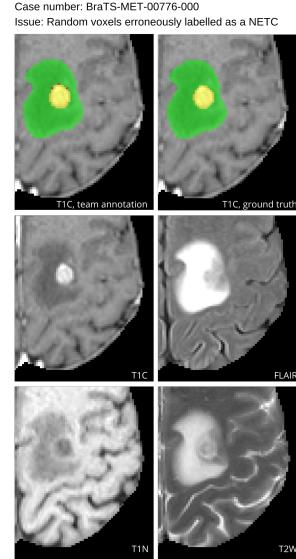
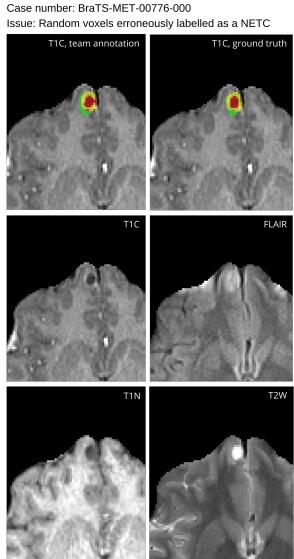


Figure 19: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

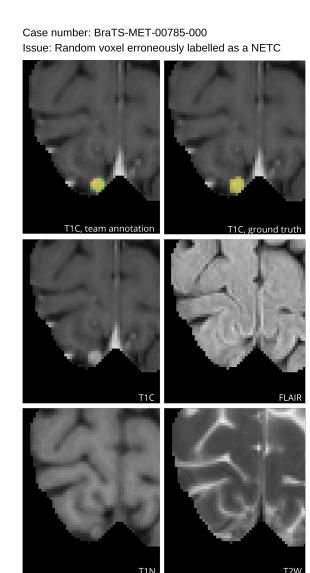
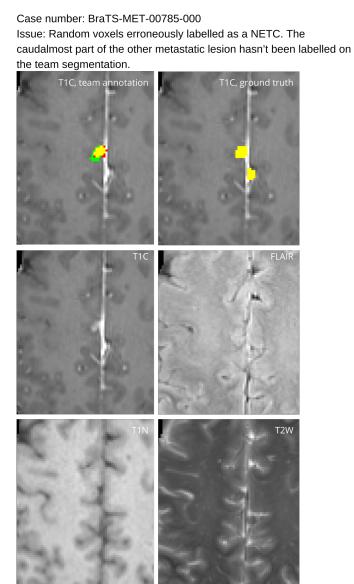
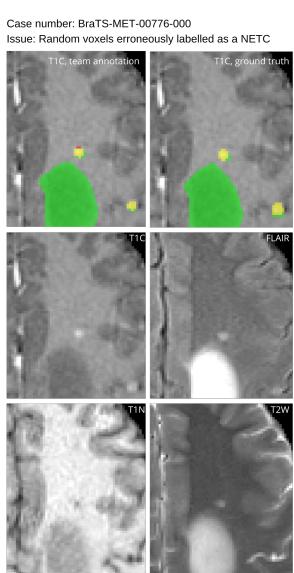


Figure 20: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

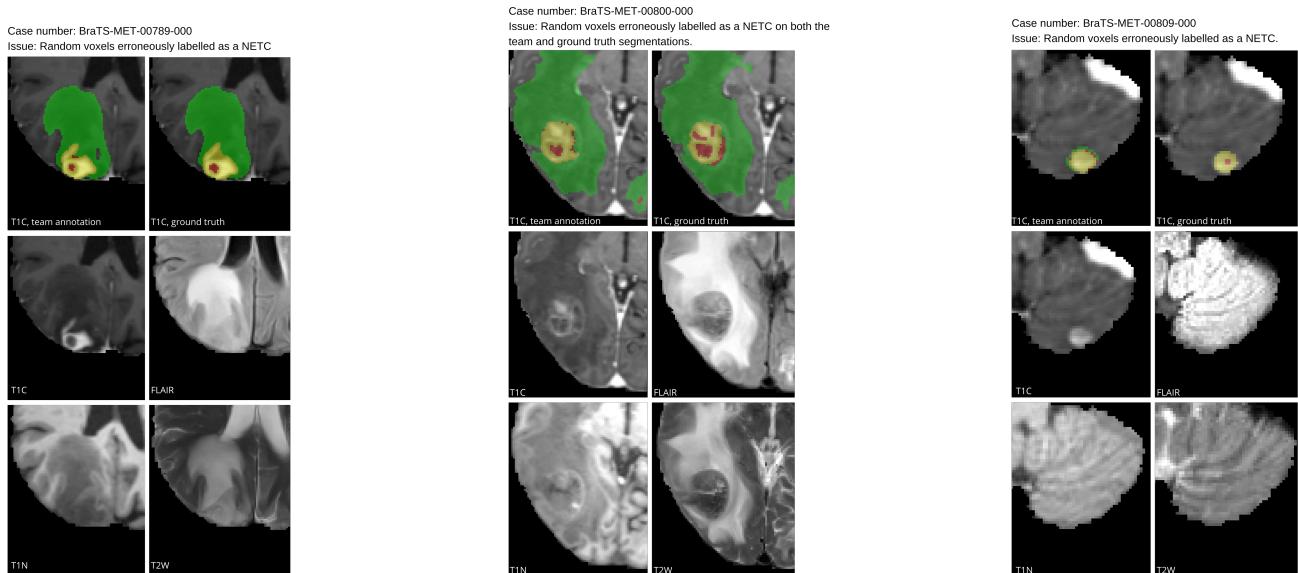


Figure 21: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

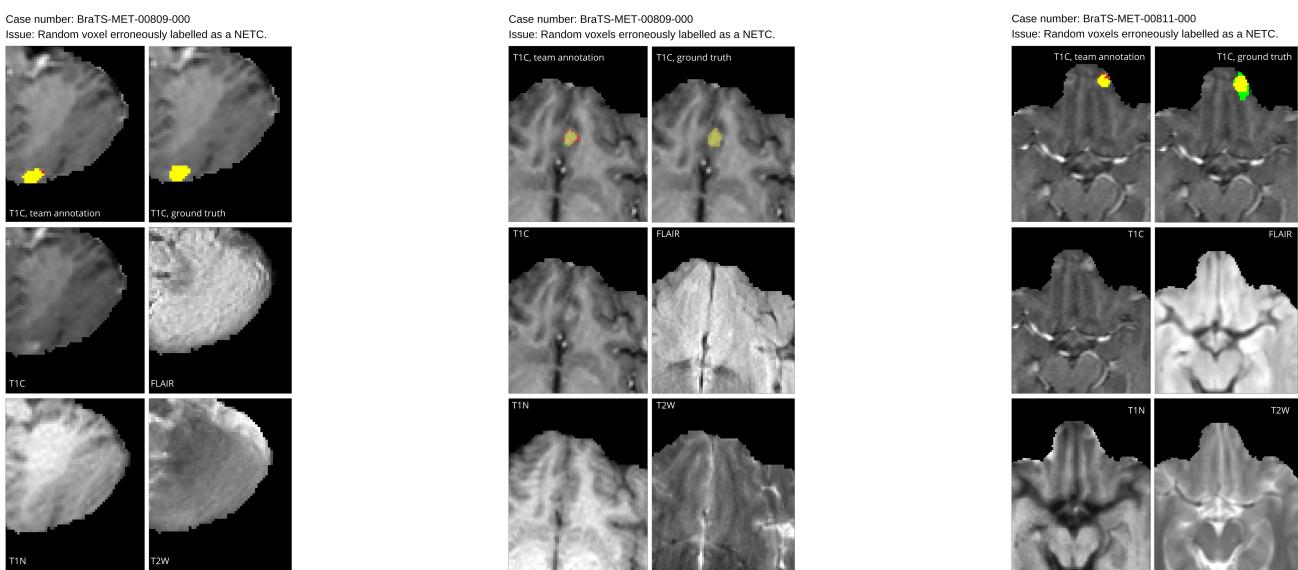
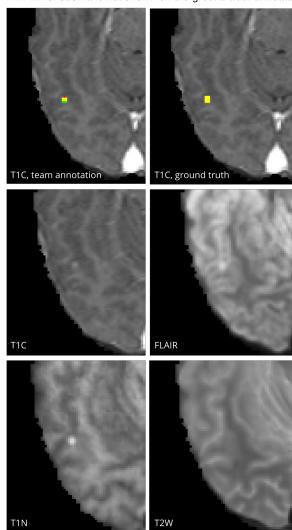
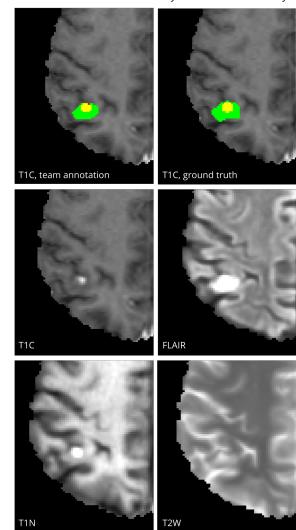


Figure 22: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

Case number: BraTS-MET-00811-000  
 Issue: Random voxel erroneously labelled as a NETC by the team.  
 Thin rim of edema is not shown on the ground truth annotation.



Case number: BraTS-MET-00811-000  
 Issue: Random voxel erroneously labelled as a NETC by the team.



Case number: BraTS-MET-00814-000  
 Issue: Random voxels erroneously labelled as a NETC.

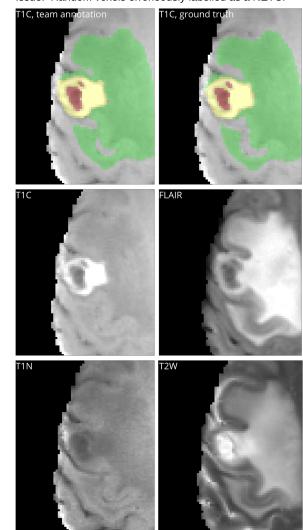
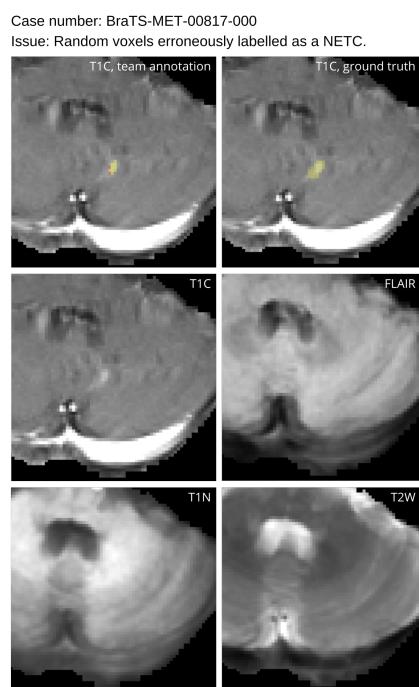


Figure 23: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core



Case number: BraTS-MET-00817-000  
 Issue: Random voxels erroneously labelled as a necrotic core

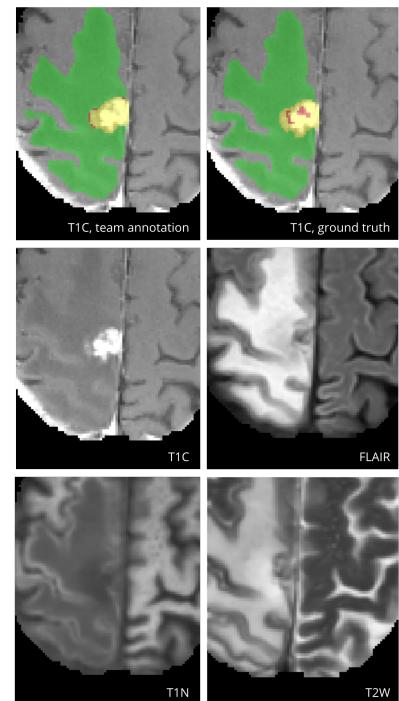


Figure 24: Supplementary: Examples of Random Voxels Predicted as Non-enhancing tumor core

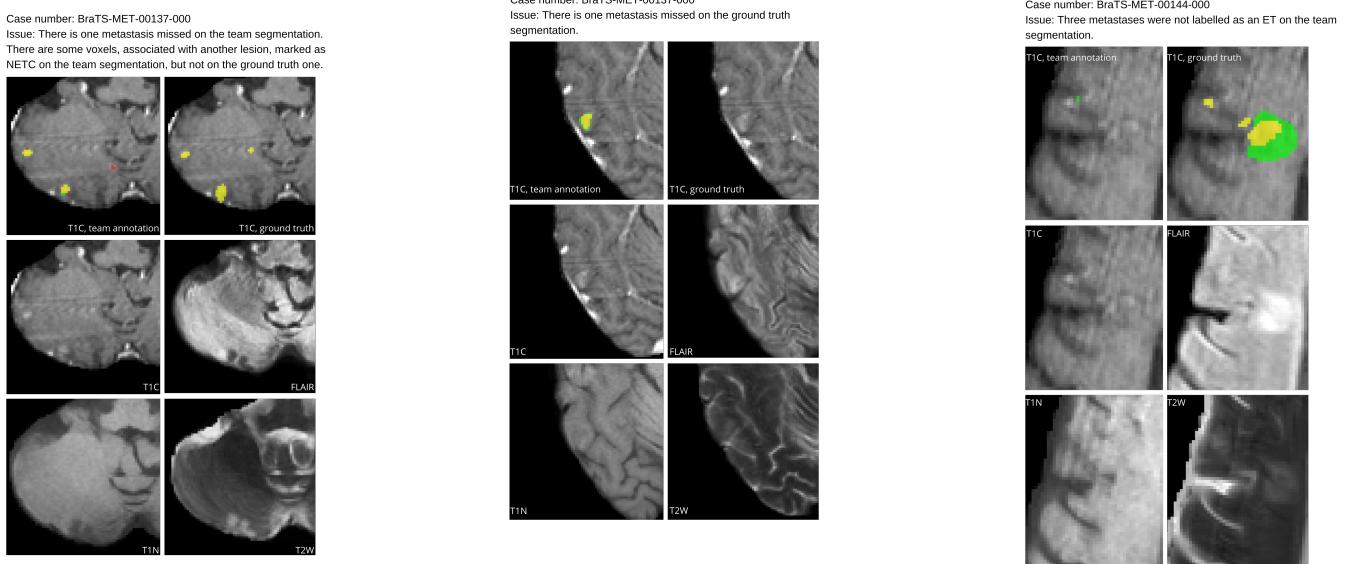


Figure 25: Supplementary: Pitfall Cases

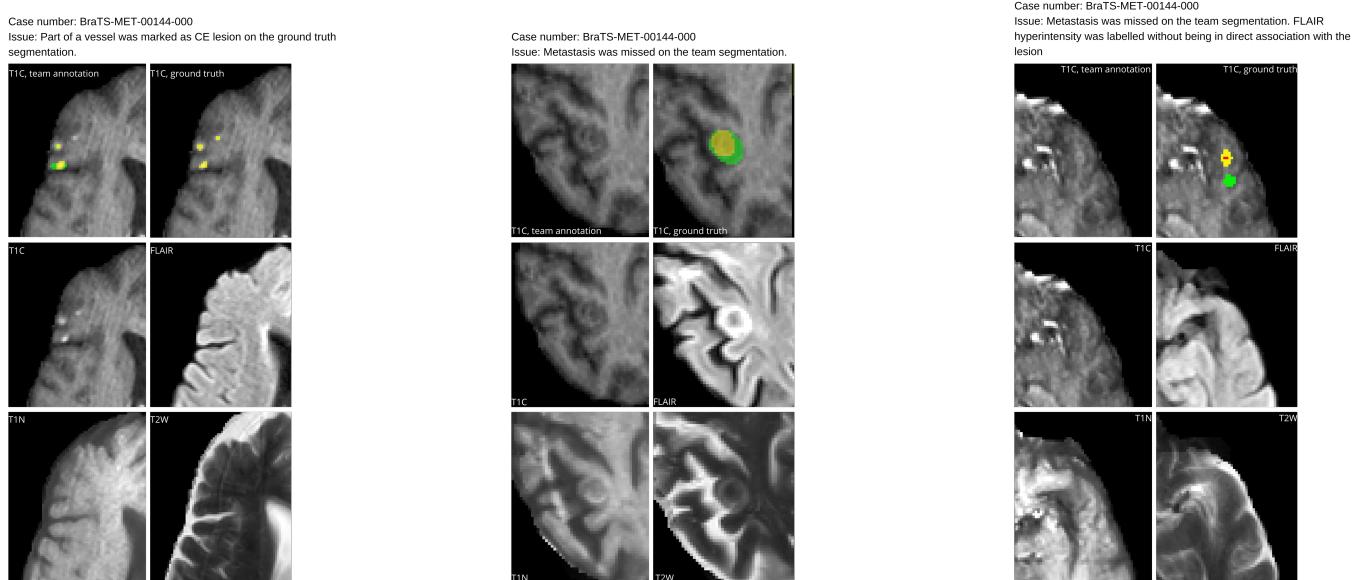
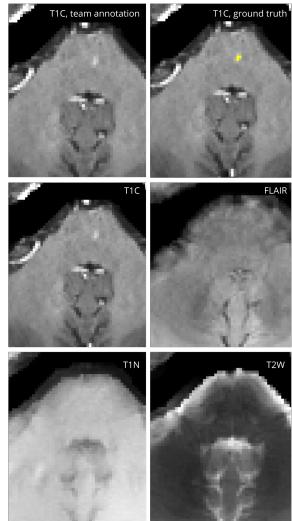
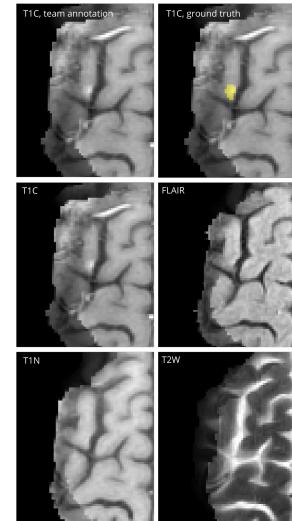


Figure 26: Supplementary: Pitfall Cases

Case number: BraTS-MET-00147-000  
Issue: Metastasis was missed on the team segmentation.



Case number: BraTS-MET-00148-000  
Issue: Metastasis was missed on the team segmentation.



Case number: BraTS-MET-00148-000  
Issue: Metastasis was missed on the team segmentation.

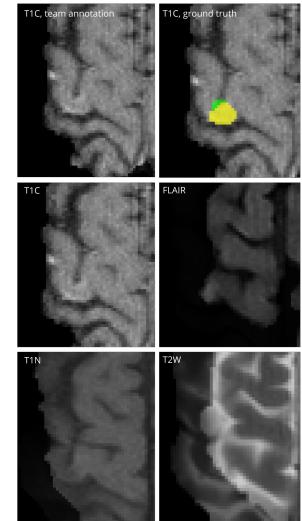
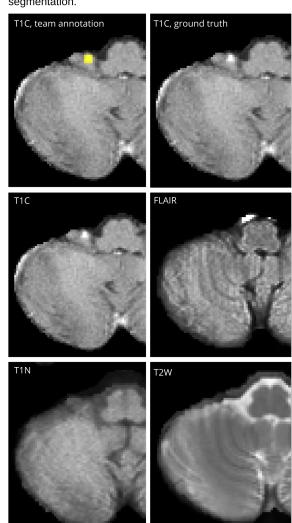
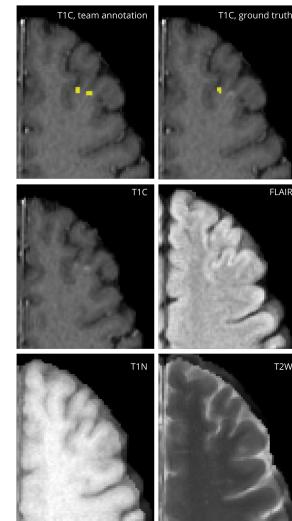


Figure 27: Supplementary: Pitfall Cases

Case number: BraTS-MET-00152-000  
Issue: A vessel was erroneously marked as a small met on the team segmentation.



Case number: BraTS-MET-00162-000  
Issue: Metastasis was missed on the ground truth segmentation.



Case number: BraTS-MET-00162-000  
Issue: Metastasis was missed on the ground truth and marked as NETC on the team segmentation.

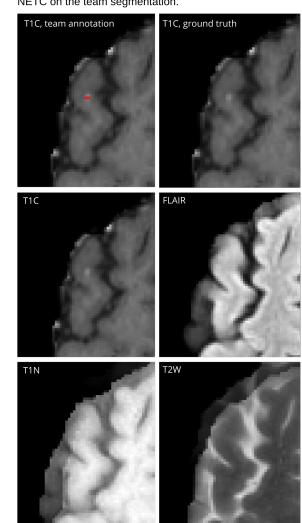


Figure 28: Supplementary: Pitfall Cases

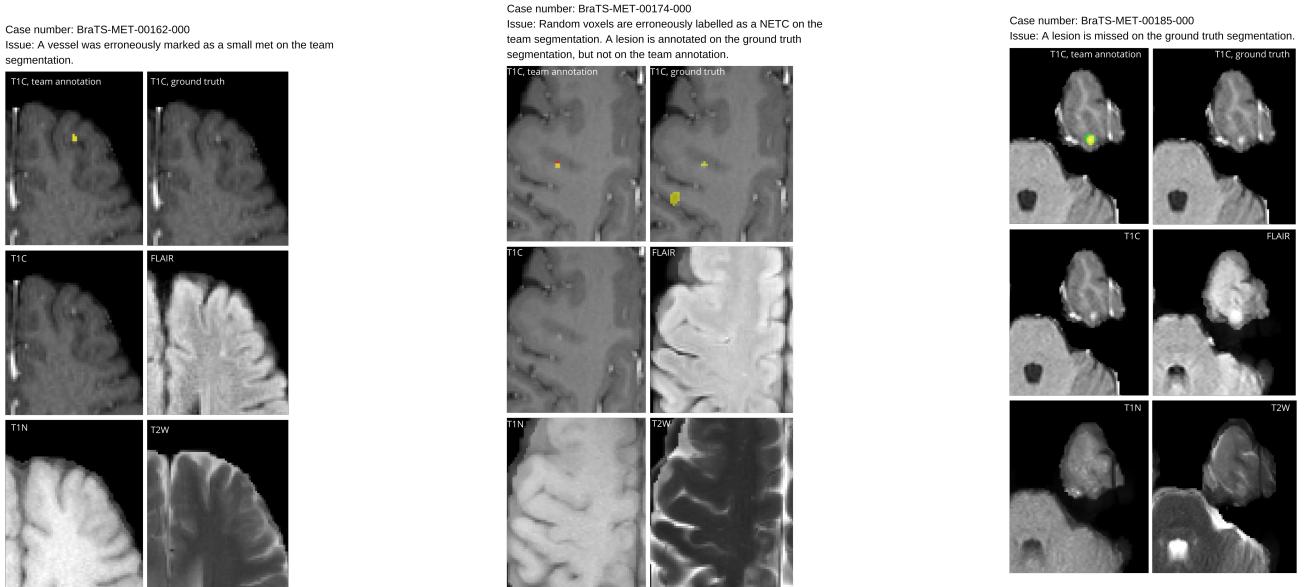


Figure 29: Supplementary: Pitfall Cases

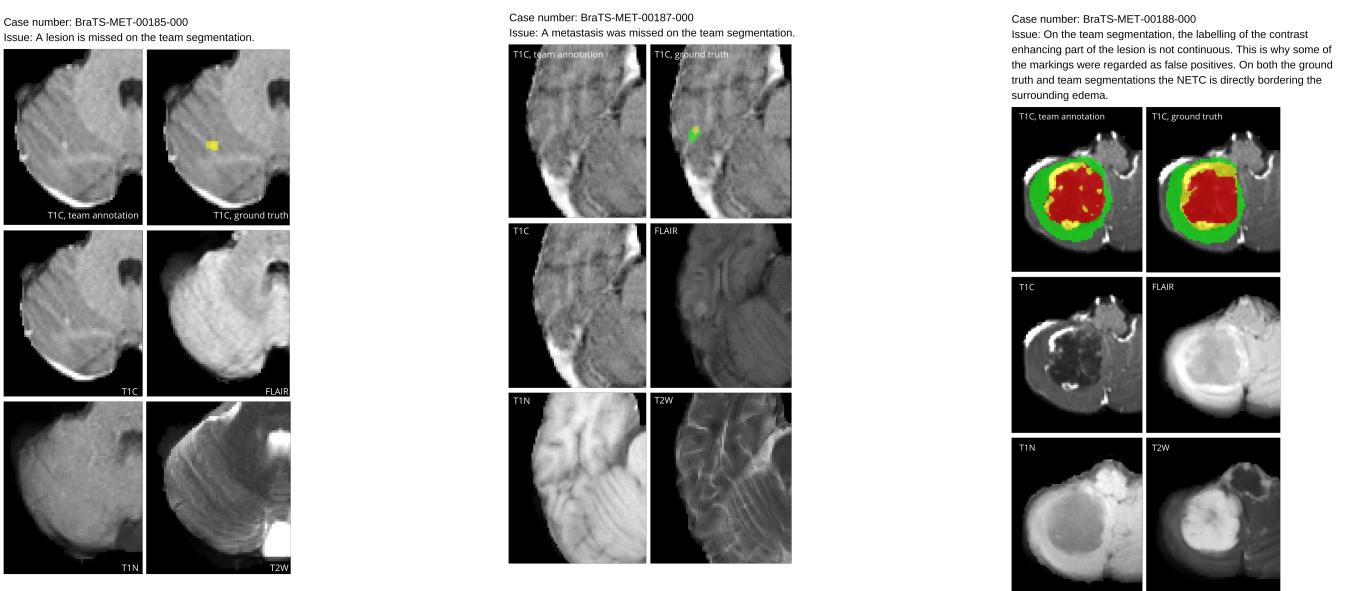
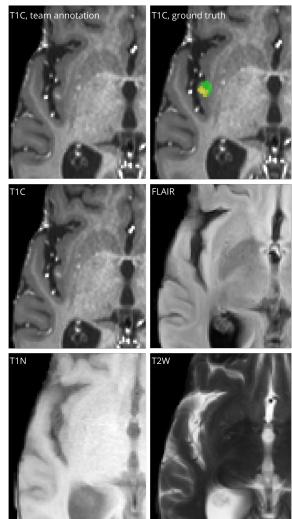
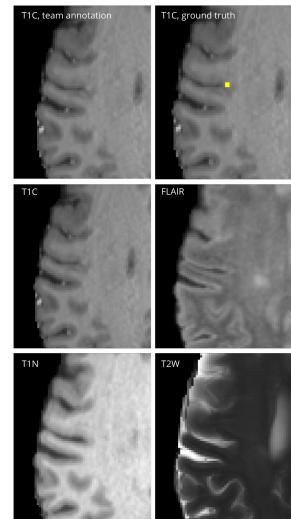


Figure 30: Supplementary: Pitfall Cases

Case number: BraTS-MET-00188-000  
Issue: There was a metastatic lesion missed on the team segmentation.



Case number: BraTS-MET-00191-000  
Issue: There was a metastatic lesion missed on the team segmentation.



Case number: BraTS-MET-00191-000  
Issue: There was a metastatic lesion missed on the team segmentation.

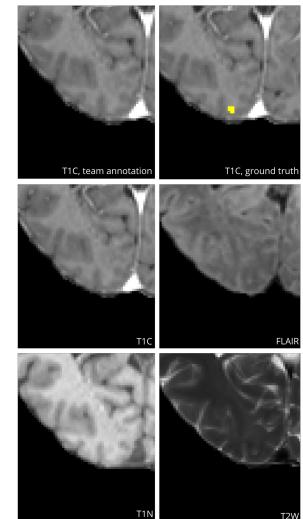
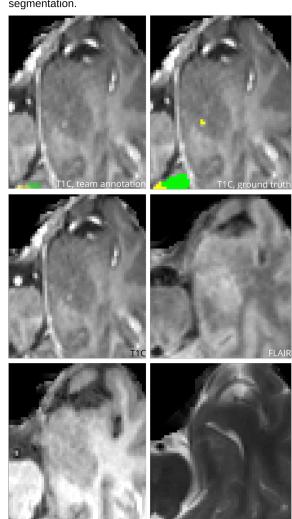
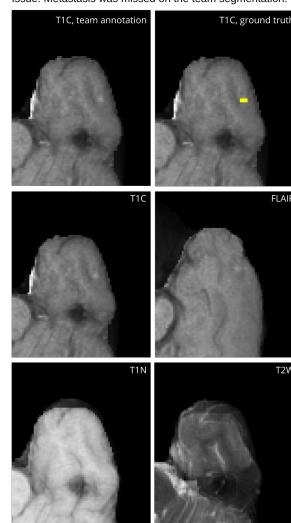


Figure 31: Supplementary: Pitfall Cases

Case number: BraTS-MET-00191-000  
Issue: There was a metastatic lesion missed on the team segmentation.



Case number: BraTS-MET-00197-000  
Issue: Metastasis was missed on the team segmentation.



Case number: BraTS-MET-00197-000  
Issue: Metastasis was missed on the team segmentation.

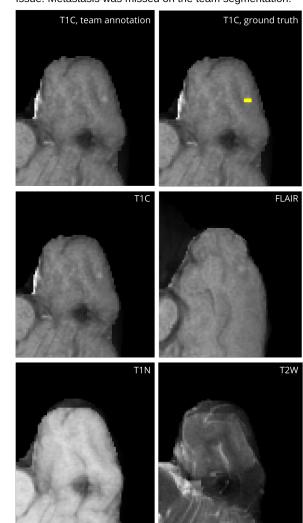
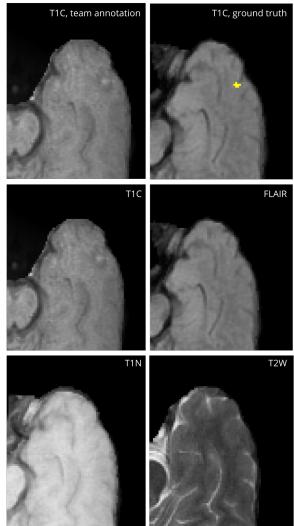
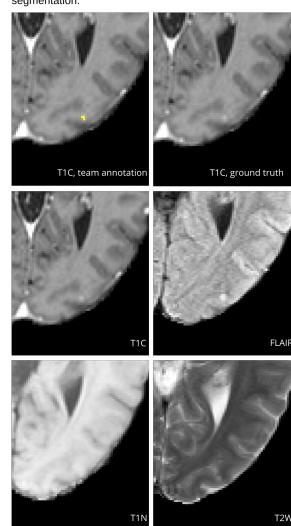


Figure 32: Supplementary: Pitfall Cases

Case number: BraTS-MET-00197-000  
Issue: Metastasis was missed on the team segmentation.



Case number: BraTS-MET-00199-000  
Issue: There was a metastatic lesion missed on the ground truth segmentation.



Case number: BraTS-MET-00203-000  
Issue: There was a metastatic lesion missed on the team segmentation.

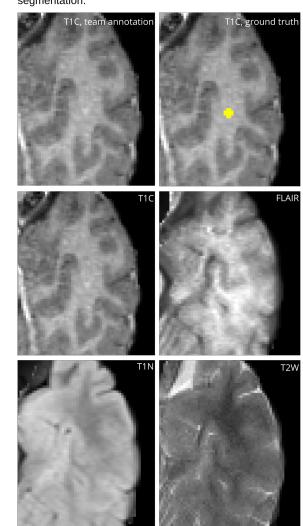
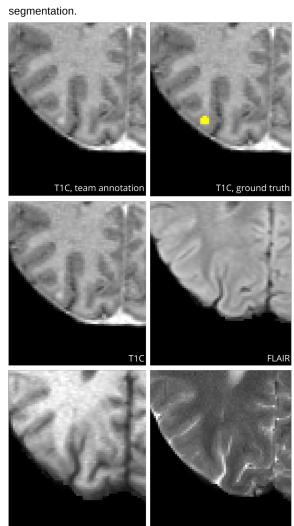
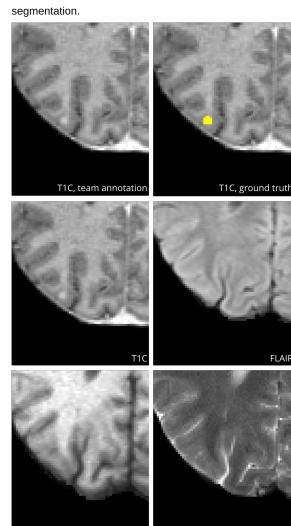


Figure 33: Supplementary: Pitfall Cases

Case number: BraTS-MET-00203-000  
Issue: There was a metastatic lesion missed on the team segmentation.



Case number: BraTS-MET-00203-000  
Issue: There was a metastatic lesion missed on the team segmentation.



Case number: BraTS-MET-00209-000  
Issue: A lesion is missed on the team segmentation.

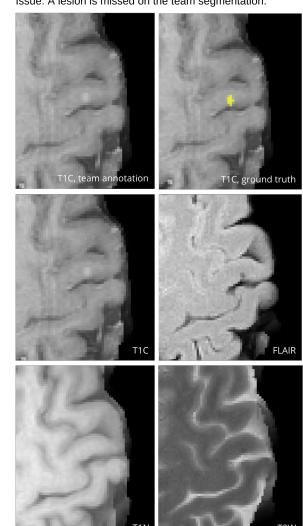
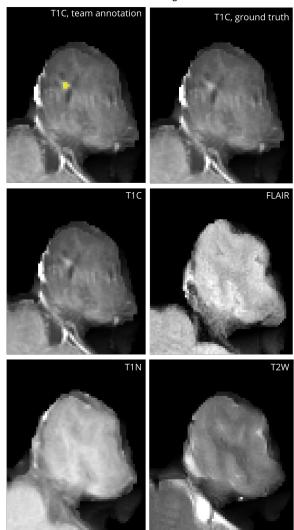
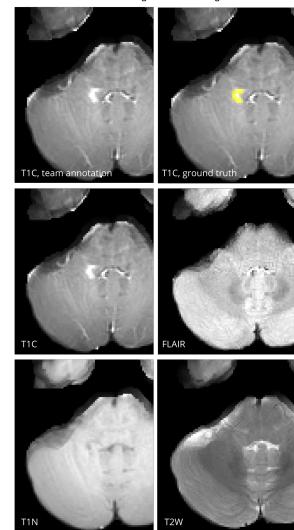


Figure 34: Supplementary: Pitfall Cases

Case number: BraTS-MET-00209-000  
Issue: What looks to be continuous with a vessel has been marked as a metastatic lesion on the team segmentation .



Case number: BraTS-MET-00209-000  
Issue: What looks to be continuous with a vessel has been marked as a metastatic lesion on the ground truth segmentation .



Case number: BraTS-MET-00213-000  
Issue: What looks like an M2 aneurysm has been marked as a met on the team segmentation.

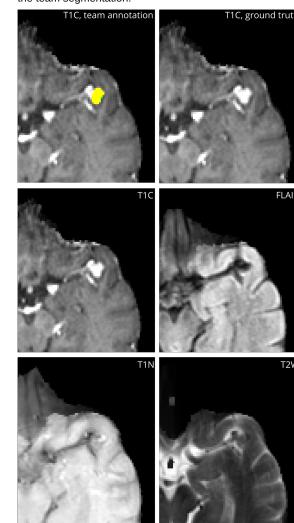
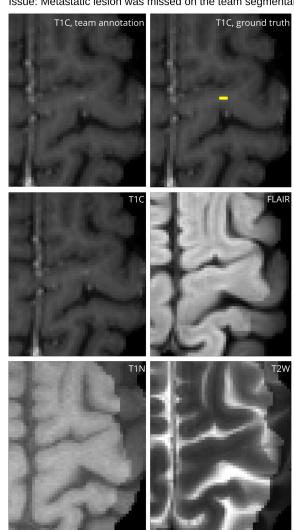
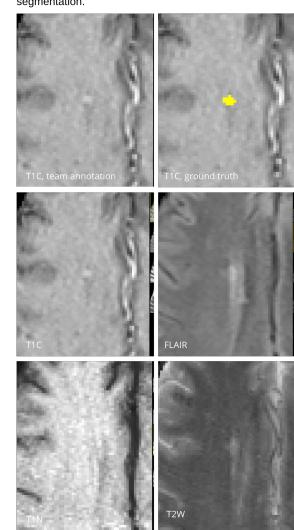


Figure 35: Supplementary: Pitfall Cases

Case number: BraTS-MET-00216-000  
Issue: Metastatic lesion was missed on the team segmentation.



Case number: BraTS-MET-00221-000  
Issue: A vessel was marked as a met on the ground truth segmentation.



Case number: BraTS-MET-00239-000  
Issue: A metastasis was missed on the team segmentation.

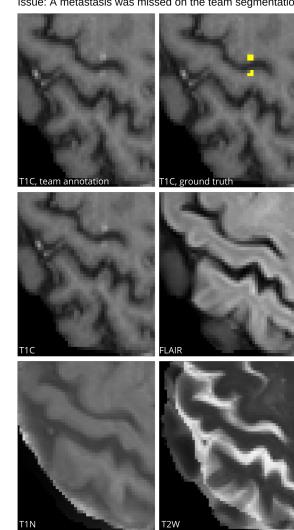
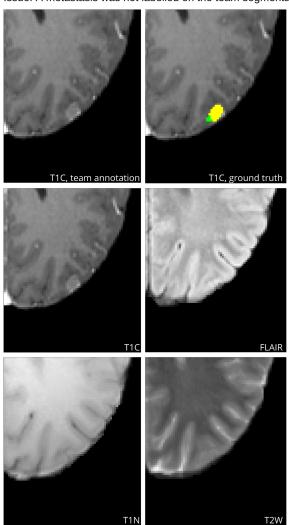
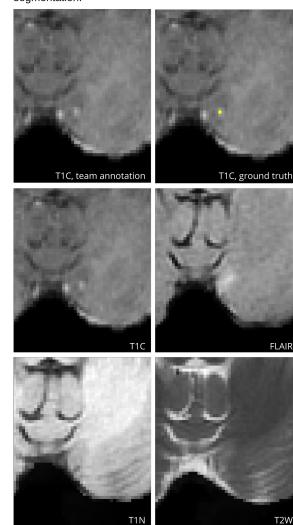


Figure 36: Supplementary: Pitfall Cases

Case number: BraTS-MET-00252-000  
Issue: A metastasis was not labelled on the team segmentation.



Case number: BraTS-MET-00276-000  
Issue: A vessel was marked as a small met on the team segmentation.



Case number: BraTS-MET-00776-000  
Issue: A small metastasis was marked as SNFH.

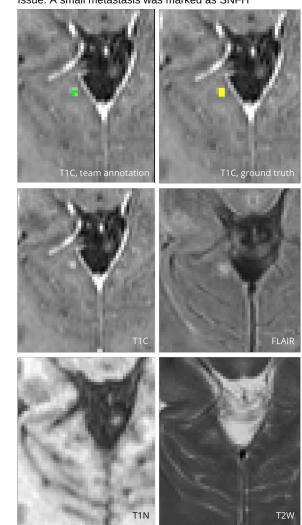
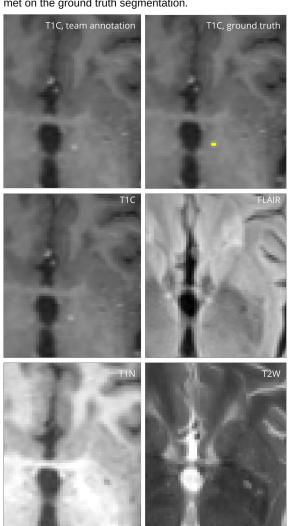
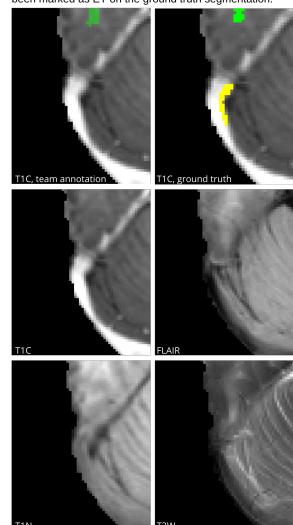


Figure 37: Supplementary: Pitfall Cases

Case number: BraTS-MET-00789-000  
Issue: What looks like a part of a vessel has been marked as a small met on the ground truth segmentation.



Case number: BraTS-MET-00800-000  
Issue: Part of the signal from transverse-sigmoid sinus junction has been marked as ET on the ground truth segmentation.



Case number: BraTS-MET-00809-000  
Issues:  
-One part of a metastatic lesion is labelled as edema on the team annotation (A) and another part of it is totally missed (B).  
-On the ground truth image, some of the necrotic core is labelled in a way that suggests it is not contained within the ET part of the tumor (A). On another slice a couple of voxels within the lesion are not assigned neither to NETC, nor to ET (B).

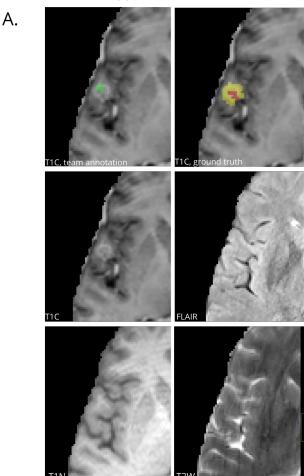


Figure 38: Supplementary: Pitfall Cases

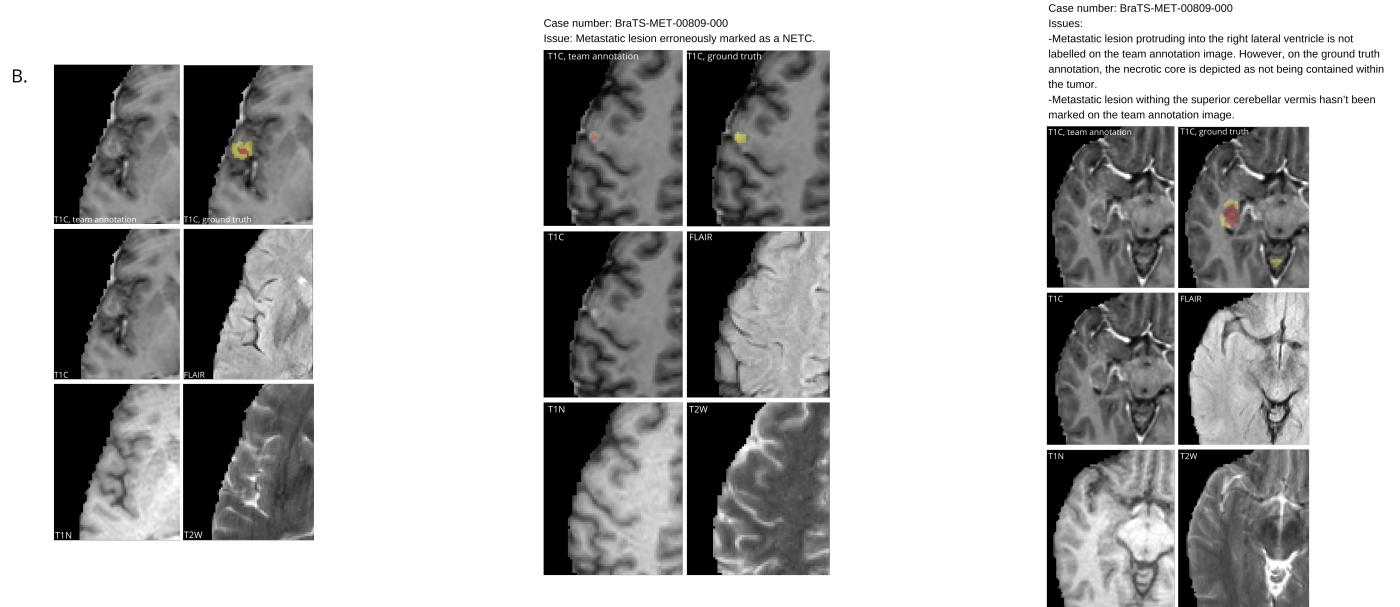


Figure 39: Supplementary: Pitfall Cases

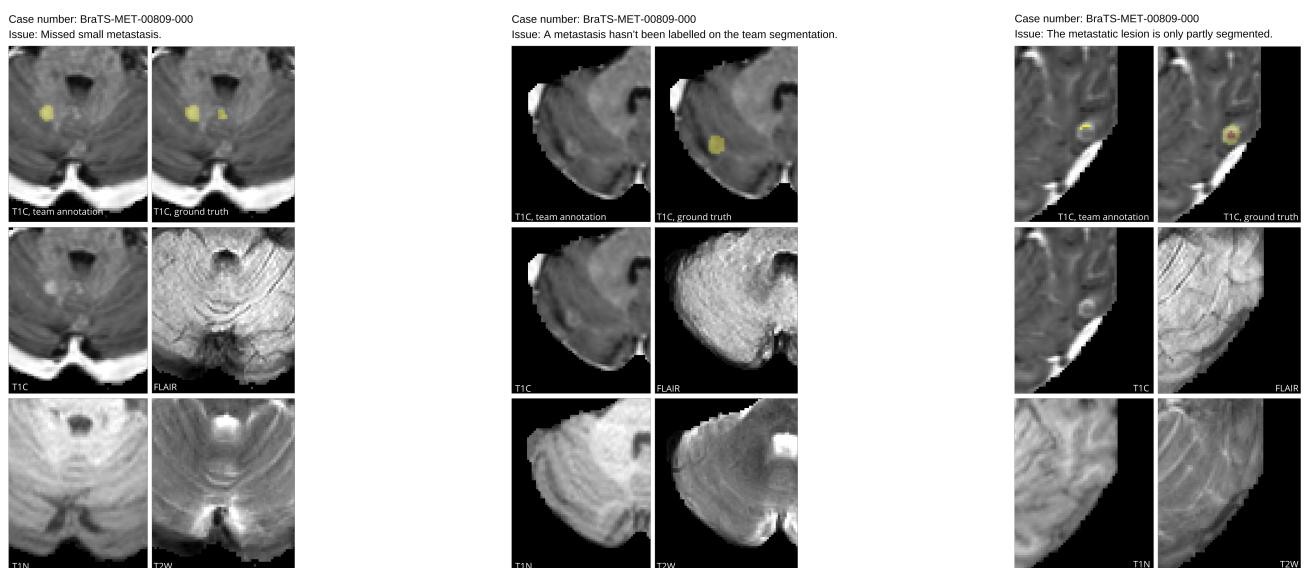
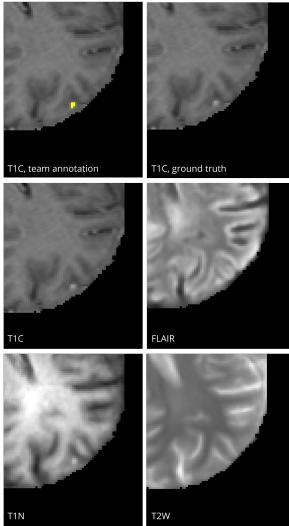
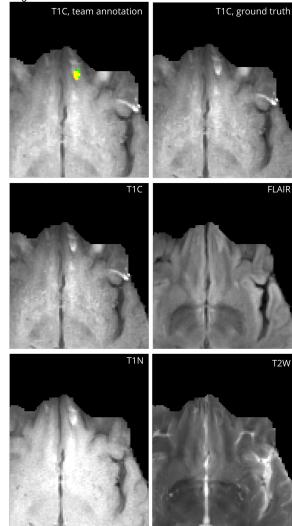


Figure 40: Supplementary: Pitfall Cases

Case number: BraTS-MET-00811-000  
Issue: A small peripheral metastatic lesion was not labelled on the ground truth segmentation.



Case number: BraTS-MET-00814-000  
Issue: Metastatic lesion was missed on the ground truth segmentation.



Case number: BraTS-MET-00817-000  
Issue: A metastatic lesion was not labelled.

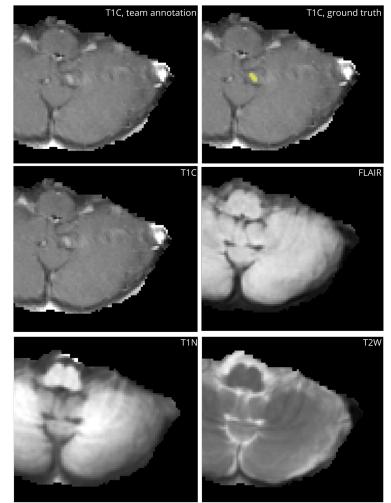
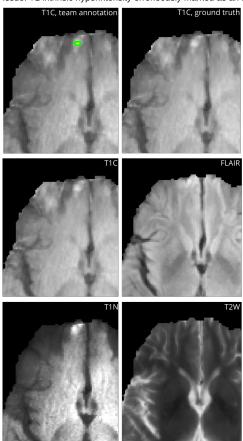
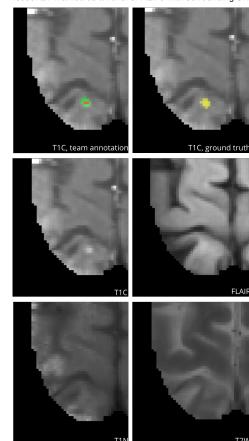


Figure 41: Supplementary: Pitfall Cases

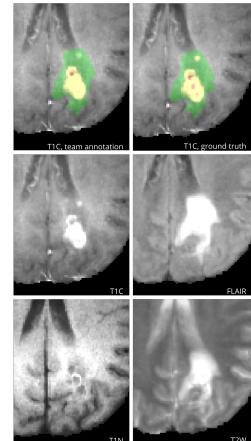
Case number: BraTS-MET-00819-000  
Issue: T1 intrinsic hyperintensity erroneously marked as an ET.



Case number: BraTS-MET-00815-000  
Issue: ET marked as a voxel of NETC with surrounding SNTH.



Case number: BraTS-MET-00822-000  
Issue: Metastatic lesion labelled as a NETC



Case number: BraTS-MET-00800-000  
Issue: The caudalmost part of the tumor hasn't been annotated on the ground truth segmentation. On the team segmentation, the ET part of the tumor has been erroneously depicted as incorporating some of the transverse sinus. Some random voxels have been incorrectly labelled as ET.

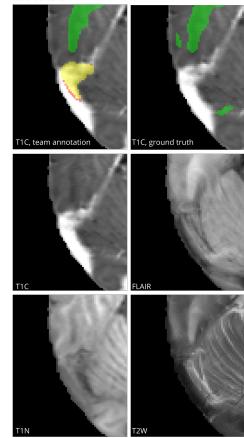


Figure 42: Supplementary: Pitfall Cases