**Medical Expenditure Panel Survey Medical Provider Component (MEPS MPC)**

Option E Summary Report

Deliverable OT-E2

Version 1.0

June 2024

Prepared for

**Agency for Healthcare Research and Quality**

AHRQ, Center for Financing,

Access & Cost Trends

5600 Fishers Lane

Rockville, MD 20857 USA

Prepared by

RTI International

3040 Cornwallis Road

PO Box 12194

Research Triangle Park, NC 27709-2194

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Author/Title | Date | Comments |
| 1.0 | Multiple Authors | 6/21/24 | Created V1.0 |

**Contents**

**Section Page**

[1. Introduction 1](#_Toc169863981)

[2. Hospital PA Updates 1](#_Toc169863982)

[3. SBD Updates: Named Entity Recognition 1](#_Toc169863983)

[4. Improving the PDF Processing Pipeline 4](#_Toc169863984)

[5. Next Steps 5](#_Toc169863986)

List of Exhibits

**Exhibit Page**

1. Named Entity Recognition for Identifying SBDs From Medical Records 3

# **Introduction**

The MEPS MPC Design Option E was awarded April 1, 2023 to expand the OBD auto-highlighting tool to the Hospital provider type. In the early months the tool’s use with both medical records (MR) and patient account (PA) records was the focus of work. However, with AHRQ’s approval, the focus pivoted to a singular objective of identifying potential Separately Billing Doctors (SBDs) once discussions began with a vendor for processing PA. This summary report documents progress toward both objectives.

# **Hospital PA Updates**

A set of Hospital PA files were processed using the OBD auto-highlighting rules to understand the extent to which this process would work for Hospital PA files. Feedback provided on these files resulted in an improvement to remove cases where dollar values were erroneously highlighted pink. For an additional improvement, an AF packet detection method was added to the processing pipeline. In the event the auto-highlighting process identifies a certain indicator symbol on a page, it will know not to highlight on that page. Both improvements reduce the number of unnecessary highlights applied by the tool, no longer requiring the abstractor or reviewer to correct these instances.

The following feedback items were deferred for resolving at a later time in favor of exploring novel approaches such as the SBD NER methods discussed above and exploring other vendor capabilities.

* + - * Discussed creating a custom Hospital PA rule for cases where individual charges for inpatient are highlighted, instead of the total charge.
      * Discussed revisiting the highlighting rules to increase the highlighting since there appears to be a general “under highlighting” occurring (by design). The sense is to revisit this and shore up the rules to highlight more while balancing this goal with not “over highlighting”.

# **SBD Updates: Named Entity Recognition**

The team investigated using Named Entity Recognition (NER) to identify potential Separately Billed Doctors (SBDs) from Hospital Medical Records (MRs). NER is a method in natural language processing (NLP) that involves identifying and classifying entities within a text into predefined categories such as names of people, organizations, locations, dates, and more. Recent innovations in NER[[1]](#footnote-2) models mean that the team can expand the standard set of entity types (e.g. Person) to domain-specific entities (e.g. Medical Professional), without having to collect new domain-specific data and train a new model for new entity types.

The team can use this approach to automatically identify names that could represent SBDs from within a document by building an NER system to identify entities of that type. Since the selected NER models do not have a concept of “separately billed doctors” as an entity type, and those models were trained on different data, the NER model can be asked to predict a generic but relevant entity type (e.g. *Medical Professional)* to identify possible SBDs, which can then expedite human review.

In addition to identifying potential SBD names, the NER model is used to also predict patient names. Because an entity can only be classified as one type, this helps improve the accuracy of the potential SBD detection by allowing us to filter out names that are predicted to be patients and not SBDs.

After entities are identified by the NER model, their location in the text[[2]](#footnote-3) is used to also capture the bounding box (i.e. spatial location in the PDF) of that entity from within the PDF box. With this information, a “context image” of the entity can be captured from within the document. This contextual image allows an abstractor or other reviewer to efficiently use the surrounding contextual information to help determine if that entity is an SBD to minimize scrolling through the underlying PDF. For a diagram of this newly developed NER process applied to SBDs, see ***Exhibit 1***.

**Exhibit 1: Named Entity Recognition for Identifying SBDs From Medical Records**

|  |
| --- |
|  |
| *Processing Pipeline for NER to identify SBDs from Medical Records* |

In the future, this system can be expanded for use on other entity types or document types. One possible direction would be exploring the identification of provider group practices from within MRs, which can then be used to verify SBDs that might not be mentioned explicitly by an individual doctor’s name. In addition, the NER models can be fine-tuned to increase performance on specific data and entity types. This would likely result in improved performance in two ways: (1) adapting the model to the unique spacing and layout of text from PDF medical records and (2) training it to more specifically identify entity types of interest for abstraction.

In order to fully launch this proof-of-concept NER solution, an updated ATO approval will be required to use RTI Merge. RTI Merge is a secure, cloud-based data science platform which would require access to the production data files. RTI maintains the Azure-based hybrid environments that are certified at the FIPS Low and Moderate impact ratings meeting the requirements of NIST SP 800-53 Rev. 4. The underlying cloud infrastructure for RTI Merge is running on a FedRAMP certified cloud provider. The ATO revision is currently being developed.

1. **Improving the PDF Processing Pipeline**

As a byproduct of the NER proof-of-concept, the team explored potential improvements to the existing PDF processing pipeline. These improvements were incorporated into the proof-of-concept for NER but are not yet implemented in production for use cases like OBD auto-highlighting.

The improvements explored during this phase of work involve:

* Detecting whether Optical Character Recognition (OCR) is necessary and only running it on the PDF when required.
* Improving the OCR accuracy when it is required for image-based documents.
* Exporting all processed PDFs as machine-readable so that text can be highlighted and copied by abstractors.

The current PDF processing pipeline runs OCR on all documents. This is necessary for documents that are image-based (scans, faxes, etc.); however, some PDFs are machine-readable by default and contain the underlying text stored as data within the PDF. This provides the benefit of directly extracting the text data from the PDF document. In comparison, OCR takes additional processing time and can introduce errors in the extracted text.

In local tests on mock MR PDFs, OCR on 15 documents (73 pages) took 2 seconds per page, while directly extracting text on those same machine-readable documents took 14 milliseconds per page—about 138x faster, on average. In addition to the performance improvements, there is no error introduced through text extraction for documents that are machine-readable – which means higher quality data for the downstream use cases like NER or auto-highlighting.

When documents do require OCR, the accuracy of the OCR process was improved by dynamically resizing the input document to a size suitable for OCR. This process uses the size of the PDF to determine the right resolution for optimal OCR results.

The team also developed a processing step that exports all PDFs as machine-readable. For already machine-readable PDFs, there is no resulting change in the output – with the exception of highlighting SBDs with NER for that use case. For documents requiring OCR, this means overlaying the OCR text onto the document in the locations that text has been recognized. The end result is a PDF where text can be selected or highlighted when viewing the PDF. This opens downstream opportunities for increased efficiencies and any associated error-reductions by having the abstractor or reviewer copying/pasting text directly instead of manually retyping data elements into the data storage system (eANF/Blaise).

These improvements for the NER proof-of-concept work are built upon a Python framework that has been upgraded from the prior version of Python used for the OBD auto-highlighting work. Upgrading Python provides the flexibility to rapidly implement the latest state-of-the-art machine learning methods for continual improvements to auto-highlighting performance and accuracy. Additionally, in deploying to RTI Merge in the future, the team can access GPUs for fine-tuning models for improving NER and/or other future models applied to highlighting processes. As noted above, a move to RTI Merge requires a revised and approved ATO.

# **Next Steps**

The MEPS MPC Design Option E work focused on innovations pertaining to auto-highlighting for both MR and PA records as well as exploring the use of NER for SBD identification. Additional efficiencies and improvements related to OCR processing were identified and applied to the NER proof-of-concept work. Following this phase of work, next steps for future work (as resources are available) focus on the following three areas:

* updating existing OBD auto-highlighting processes
* expanding NER capabilities
* integration with existing data entry systems

The NER proof-of-concept work for SBD identification yielded two secondary outcomes, resulting in efficiencies and accuracy by inputting machine-readable PDFs without the use of OCR (when applicable) and outputting PDFs in machine-readable format. Moving forward, the OBD auto-highlighting processes could be updated to read in machine-readable PDFs and only apply OCR on an as-needed basis. Similarly, an update could be applied to output all highlighted PDFs as machine-readable enabling the abstractor to highlight and copy/paste text into data entry systems without a need for manually typing values identified in the PDF.

The existing NER proof-of-concept work was limited to a set of mock MR PDFs. Moving forward from this phase of work, the team could use production MR PDFs to fine-tune the NER model to improve the accuracy identifying SBDs. A document level text processing approach will be implemented to supply the optimal amount of information to the NER model to result in further accuracy gains. In addition, NER capabilities could be expanded to identify new entities of interest such as provider groups and organizations, using production data to fine-tune the model and improve accuracy. NER could be expanded to new document types beyond MR files where the OBD auto-highlighting may not be accurately capturing a value that NER could potentially successfully identify. As highlighted in the NER summary above, implementing NER would necessitate a move to RTI Merge which requires a revised and approved ATO.

The current PDF processing pipeline involves extracting text data from each PDF to then perform NER or apply OBD auto-highlighting rules resulting in a final highlighted document for the abstractor to review. Data is then manually typed into data storage systems (eANF/Blaise) by the abstractor or reviewer. The team will explore the option for a data integration process to export these highlighted fields and their corresponding values into the data storage system, removing the need for a manual data entry step.

1. Zaratiana, Urchade, Nadi Tomeh, Pierre Holat, and Thierry Charnois. “GLiNER: Generalist Model for Named Entity Recognition Using Bidirectional Transformer.” arXiv, November 14, 2023. <https://doi.org/10.48550/arXiv.2311.08526>. [↑](#footnote-ref-2)
2. The *text location* of an entity is their start character and end character from within the full text extracted from the PDF. For example, in the sentence “*Paul* picked pickles”, the entity *Paul* spans from characters 0 to 3. [↑](#footnote-ref-3)