Methods

Between September 2015 and December 2017, all HL7 CDA records were processed manually by Westat staff. Between December 2017 and February 2018, Westat staff under the MCR project sought to improve the established architecture and data flow for processing the HL7 CDA records by consulting with Westat data scientists. The goals were to:

- Develop methods that required less manual processing
- Develop a process that could handle the large amounts of data that the MCR had received

The data scientists analyzed the architecture and processes in place in order to develop candidate processing steps to streamline and refactor. After evaluation, they created an automated data processing pipeline that would require no staff intervention.

- The automatic pipeline uses three discrete services: receipt, storage, and eMaRC.
- All services run as Windows services on a project-specific server.
- Westat implemented the pipeline for the MCR in February 2018.

The automatic pipeline uses three discrete services:

- **Receipt**: Invokes eMaRC Plus to process the files.
- **Storage**: Writes documents to a temporary location.
- **Process**: Runs patient-level match against MCR database; sends non-matches to eMaRC.

At the time of manual processing, a patient-level match was performed using a SAS program. In the automatic processing pipeline, the data scientists analyzed the architecture and processes in place in order to develop candidate processing steps to streamline and refactor. After evaluation, they created an automated data processing pipeline that would require no staff intervention.

- The automatic pipeline uses three discrete services: receipt, storage, and eMaRC.
- All services run as Windows services on a project-specific server.
- Westat implemented the pipeline for the MCR in February 2018.

Results

- **Manual Process**
  - 26,620 Files prepared for eMaRC
  - 65% Matches (non-matches sent to eMaRC)
  - 35% Non-matches (Consisting of 2,881 new MCR cases)

- **Automatic Pipeline**
  - 91,185 Files prepared for eMaRC
  - 66% Matches (non-matches sent to eMaRC)
  - 34% Non-matches (Consisting of 3,546 new MCR cases)

**Annual Savings**

- 3.5X More Files, in less than half the time
- 3.5X More staff time will be spent managing the import of files into eMaRC Plus, exporting the abstracts out of eMaRC Plus, and running a match SAS program against the abstracts to identify non-matched cases.

No more staff time will be spent managing the import of files into eMaRC Plus, exporting the abstracts out of eMaRC Plus, and running a match SAS program against the abstracts to identify non-matched cases.

Matching cases against the MCR database prior to import into eMaRC Plus results in fewer files that need to be processed and less time spent by MCR on processing files that may not yield useful information.

- The pipeline processes more tasks manually, at a faster rate, saving 3-4 hours of staff time per week.
- Polling for new HL7 CDA documents generally completes in a few minutes, and the function can receive multiple files per second.
- The storage function can complete tasks in as fast as a few hundred milliseconds, or as long as several minutes for extraordinarily large files.
- HL7 CDA documents are no longer stored on a study-specific file storage system. This provides more available storage space for the project, which is a valuable resource.

Next Steps

- Incorporate exporting abstracts from eMaRC Plus into the automatic processing pipeline.
- Facilitate quality assurance activities. The automatic processing pipeline will expedite the time from data receipt to abstractor review (which has previously taken about 3 years).

Conclusions

It is important to evaluate current processes for data management improvements that can lead to reduced costs and increased workload efficiency. The MCR has already experienced improvements with the implementation of the automatic processing pipeline for HL7 CDA documents received for MUR. The pipeline performs more tasks than staff were completing manually, and at a faster rate, saving 3-4 hours of staff time per week.

- Polling for new HL7 CDA documents generally completes in a few minutes, and the function can receive multiple files per second.
- The storage function can complete tasks in as fast as a few hundred milliseconds, or as long as several minutes for extraordinarily large files.
- HL7 CDA documents are no longer stored on a study-specific file storage system. This provides more available storage space for the project, which is a valuable resource.

Acknowledgments
Cancer data have been provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health. We acknowledge the State of Maryland, the Maryland Cigarette Restitution Fund, and the National Program of Colorectal Cancers of the Centers for Disease Control and Prevention for the funds that support the collection and availability of cancer registry data.